

Facing the nuclear danger

The war on terrorism threatens to overshadow the greatest weapons-proliferation challenge of all — the safe management of nuclear materials in the former Soviet Union.

Saddam Hussein, Osama Bin Laden and Kim Jong-Il are each well-known for their possible ties to weapons of mass destruction, but has anyone heard of Alexander Tyulyakov? Tyulyakov was deputy director of Atomflot — the state-owned company overseeing Russia's fleet of nuclear icebreakers. In August, he was arrested by the Russian Federal Security Service and charged with trying to sell more than a kilogram of uranium and radium to nuclear smugglers.

When Tony Blair and George W. Bush meet in London this week, it must be hoped that their discussions on the threat of proliferation of weapons of mass destruction will not be confined to the status of weapons programmes in Iraq, North Korea and Iran.

Proliferation experts have argued for years that the kind of activities in which Tyulyakov was allegedly engaged present an immense danger. This was acknowledged soon after the collapse of the Soviet Union, in particular by then senator Sam Nunn (Democrat, Georgia) and Senator Richard Lugar (Republican, Indiana), who co-sponsored a 1991 law to support the adequate supervision of nuclear materials there. But it has been difficult for the United States to sustain its commitment to Nunn–Lugar programmes. And European governments have shown even less commitment — lamentably, some of them continue to see this pressing issue as something for the United States and Russia to address (see page 219).

But the problem belongs to all of us and it isn't going away. Its scope spreads far beyond the old Soviet atomic-weapons complex and the nuclear submarines' fuel cycle. Across Russia, for example,

hundreds of weather stations and navigation beacons are powered by highly radioactive strontium, which could be used in so-called 'dirty' weapons that spread such materials around without any fission reaction. And shipyards and factories house tonnes of nuclear fuel from ships and power plants. Research reactors in the former Warsaw Pact countries pose another risk. That's aside from the tonnes of biological agents thought to reside in Russian military installations, and the thousands of intact nuclear weapons still deployed or stored in the country.

The Nunn–Lugar act and other salient measures anticipated a prolonged and concerted Western effort to help safeguard these materials. But a report released this week by a coalition of 21 non-proliferation organizations suggests that the effort is in danger of stalling.

The report (see www.sgpproject.org) describes the progress to date of the Global Partnership Against the Spread of Weapons and Materials of Mass Destruction, an initiative launched in June 2002 to secure government commitments of US\$20 billion over ten years to clean up facilities, manage materials and gainfully employ weapons scientists in the former Soviet Union. But in the first year of the partnership, the report says, its 14 member governments spent only about half of the US\$2 billion needed to meet its goals.

If governments are serious about slowing weapons proliferation, there can be no higher priority than making up this shortfall. And that will only be the start of the decades-long commitment required to contain the Soviet legacy of weapons of mass destruction. ■

Local support required

The Royal Society's review of Britain's university funding system should take into account the needs of the regions.

British universities are, by most measures, doing well in their ability to perform internationally competitive research. But there's a widespread feeling that they could do even better. This week, the Royal Society has called for what it terms a "radical review" of the dual support system, whereby the universities get most of their operating costs from higher-education funding councils, and most of their research money from research councils.

The review is worthwhile, especially given the level of carping that surrounded the most recent attempt to track the performance of university departments, through the higher-education councils' Research Assessment Exercise. That exercise has played a useful role, but seems to have run its course.

So, what next? British universities have recovered from a serious crisis of confidence that reached its nadir about twenty years ago. Lately, their administrators' chief priority has been to emulate the American institutions that continue to poach faculty at a fair clip.

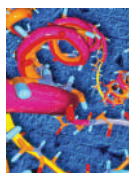
This has involved pestering alumni for cash contributions. Unfortunately, British university graduates with £1,000 to spare seem more inclined to buy a season ticket to watch Arsenal than to bail out their Alma Mater — even if the former isn't tax-deductible. Attempts to build private sources of funding for British universities — or

even to establish private universities — have foundered, meanwhile, on the widespread assumption that university education (and, by extension, university research) is essentially a public function.

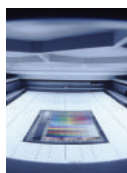
So which of the tools that built the US research university are really available to Britain? One that you don't hear much about is the ability of American states to back their own public universities to the hilt. In this way, the likes of the University of Wisconsin–Madison and the University of Arizona at Tucson have been able to build truly world-class departments that can compete for staff and research funds with longer-established institutions on both US coasts.

Despite early pledges from Tony Blair's first government, the English regions have no fiscal power to support vital institutions of their own, such as universities. Business leaders and politicians in and around Manchester, Leeds, Nottingham and Bristol, for example, would like nothing better than to build up local universities, as their counterparts in Germany and the United States are able to do. But they can't. The newly established Scottish parliament is already taking steps to boost its own university system. A similar mechanism is needed in England if the Oxford and Cambridge duopoly is to be replaced with a flexible and genuinely competitive university system. ■





Close scrutiny
National Academy
calls for better test
for prion proteins
p216



Double trouble
Academics face
massive bills for
photocopying
p217



Caught out
Germany considers
unfair fate of
whistleblowers
p218



No effect
AIDS vaccine fails
second large-scale
trial
p220

Medical journal under attack as dissenters seize AIDS platform

Declan Butler

The *BMJ* (*British Medical Journal*) is under fire from AIDS researchers over a series of publications on its website. The postings in question make up a steady stream of unreviewed articles from people who deny that HIV causes AIDS.

The dispute crystallizes the conflict between a journal's desire to experiment with open, unmoderated electronic debate and its fundamental obligation to readers to provide them with authentic information, researchers say.

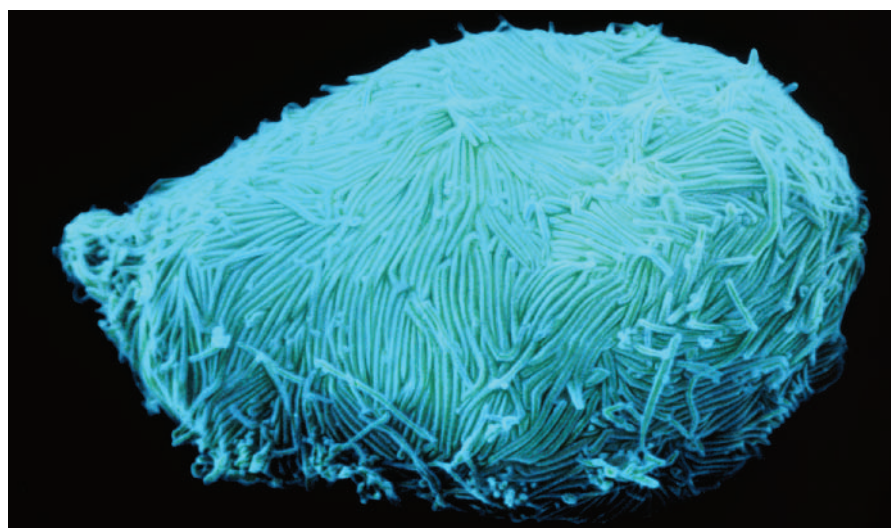
If you search the *BMJ*'s website for 'AIDS' in all articles published over the past two years, the results are surprising. One in six of the top 300 returns are written by AIDS 'revisionists'. Add in responses to these missives, and more than one-third of the returned articles centre on that debate.

The disproportionate prominence of this discussion compared with other AIDS issues is angering some researchers. They point out that almost all of the articles concerned are unreviewed 'Rapid Responses' — electronic letters to the editor that are published within 24 hours, and are screened only to exclude libel or breaches of patient confidentiality.

David Rasnick, a leading AIDS sceptic, accounts for 46 Rapid Responses published since January 2002. Rasnick, a chemist who is a visiting scholar at the University of California, Berkeley, gives his affiliation as "a member of the Presidential AIDS Advisory Panel of South Africa". Another sceptic, Eleni Papadopoulos-Eleopoulos, a biophysicist at Royal Perth Hospital in Australia, has made 33 postings.

Researchers complain that the credibility conferred on Rapid Responses by the *BMJ* label may mislead policy-makers and members of the public over the debate. "It looks to the unsuspecting that this is solid stuff," says Simon Wain-Hobson, who studies AIDS at the Pasteur Institute in Paris. Wain-Hobson is one of several AIDS researchers who argue that unfettered debate on the *BMJ* website offers revisionists an opportunity to punch well above their weight in a renowned journal.

Brian Foley, an AIDS researcher at the Los Alamos National Laboratory in New Mexico, is one of the scientists most active in respond-



The *BMJ*'s website carries postings that deny that HIV, seen here in a white blood cell, causes AIDS.

ing to the revisionists' *BMJ* postings. He criticizes the *BMJ* for allowing what he describes as the misuse of a respectable scientific journal to contribute to the dissemination of disinformation. "I do not think responding to *BMJ* posts is a worthwhile use of my time," he says.

Many scientists argue that not responding is the best policy. John Moore, an AIDS researcher at Cornell University in New York, is an example. He says that revisionists are best ignored: "It's an unwinnable debate based on faith not fact."

Richard Smith, editor of the *BMJ*, admits that a problem of unmoderated forums is that one side of an argument can "lose by attrition" to a more vocal minority. But Smith also cherishes the principle of Rapid Responses, which have solicited more than 20,000 submissions since being introduced five years ago. Its philosophy, says Smith, is based on free speech.

But such aspirations cut little ice with Moore. "The *BMJ* is wrong to grant the revisionists space in what should be considered a prestigious journal," he says. The quality threshold for publishing Rapid Responses is "obviously very low," says Smith, who admits that some people abuse the system by "pursuing vendettas or obsessions, by writing every day, or at excessive length".

But Smith feels that this is the price to pay

for the benefits of free debate. "Freedom of speech means you end up with lots of abuses," he says. "I continue to think it is the right thing to do." But whereas the *BMJ*'s editorial board has unanimously endorsed this general principle, he says that as a result of researchers' criticisms, he may raise the specific issue of AIDS revisionism with both this body and the journal's ethics committee.

Rasnick says that he "applauds the *BMJ* for having the courage and integrity to allow debate" and questions why mainstream AIDS researchers complain so much "if their position is so secure". He says that he wants to know why "over 100,000 scientists and doctors, the 140,000 papers, and the \$118 billion spent all combined have not come up with the proof that AIDS is contagious; that AIDS is sexually transmitted; that HIV causes AIDS; that anti-HIV drugs do more good than harm; or that AIDS is devastating and depopulating Africa or anywhere else." Most AIDS researchers strongly dispute these assertions.

Smith adds that in the immediate future the *BMJ* may add a visible health warning to its Rapid Responses to indicate their unreviewed nature, or may change the *BMJ*'s default online search option — currently set to include 'Articles and Rapid Responses' — to 'Articles only'.

Syngenta ends plant-research deal with Berkeley

Rex Dalton, San Diego

An industrial-academic partnership on plant research that rocked the University of California (UC), Berkeley, when it started five years ago is quietly coming to an end.

The deal, under which Syngenta, the Swiss agricultural biotechnology firm, provided \$25 million to the university's department of plant and microbial biology, will expire on 23 November, after the company declined to exercise its option to continue it. The deal was originally drawn up by Novartis, which spun off its agricultural interest in 2000 to form Syngenta.

"There is a shift in how we do discovery research," says plant biologist Simon Bright, head of technology interaction for Syngenta at Jealott's Hill, Berkshire, UK. "We are focusing on moving discoveries in the pipeline to products."

UC Berkeley officials say that they are unsurprised, but that they would have liked to renew the agreement. "It funded a lot of blue-sky research that would not otherwise have taken place," says political scientist Robert Price, UC Berkeley's associate vice-chancellor for research.

Early next year, an independent analysis of the partnership's impact is to be completed by a research team from Michigan State University (MSU) in Lansing. The \$225,000 study — run by sociologist Lawrence Busch of MSU's Institute of Food and Agricultural Standards — was commissioned by the UC Berkeley academic senate.

When the Syngenta pact was revealed five years ago, it caused a fierce argument about relationships between industry and university departments (see *Nature* 399, 5; 1999). It gave faculty and students access to dynamic technologies, such as proprietary plant sequence databases, but allowed the corporation to keep rights to discoveries.

UC Berkeley officials say that Syngenta reviewed some 375 abstracts of scientific research undertaken by faculty members and students. Preliminary discussions have been held about the corporation licensing one undisclosed discovery for development, officials say. Plant pathologist Brian Staskawicz, the UC Berkeley professor who supervised the deal, says: "I think everyone would agree the collaboration was a great success. It really was an experiment." Other faculty members say that they will await the results of the outside analysis before commenting. ■

Academy calls for improved tests to beat prion disease

Jonathan Knight, San Francisco

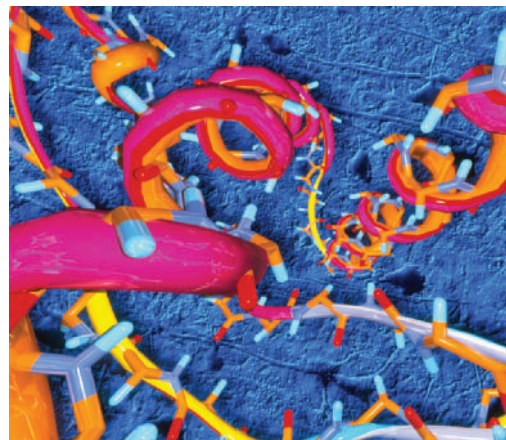
Diagnostic tests for prion diseases need a "quantum leap" in sensitivity if they are to help prevent future outbreaks, says a report from the US National Academy of Sciences.

Basic research into the nature of prions, which cause illnesses such as bovine spongiform encephalopathy (or mad cow disease), will be essential to achieve this, says the report. It was commissioned by the Department of Defense and prepared by a panel chaired by Richard Johnson, professor of neurology at Johns Hopkins University School of Medicine in Baltimore, Maryland.

"We need a fundamentally different approach," says Johnson. "Incremental improvements in existing tests are not going to do it." Prion diseases, or transmissible spongiform encephalopathies (TSEs), can only be confirmed after death and even then can be missed. A test is needed that can spot the smallest amount of infectious prion protein that can cause disease, the report says. This would require a sensitivity at least 1,000 times greater than that of current tests.

One possible approach would be to work out the three-dimensional structure of the prion protein, so that a compound could be designed to target it. This has been complicated by the insolubility of prion proteins.

Fred Cohen, a prion researcher at the University of California, San Francisco,



Knotty problem: the prion's insolubility slows research.

welcomes the report. "It was a good group of people to probe these questions," he says. Although there has been an epidemic of TSE among elk and deer in the Midwest, there is no mad cow disease in the United States, and Cohen doubts whether the political will exists to fund the work.

Congress handed a major role in prion research to the defence department last year, when it authorized the National Prion Research Program. This distributed US\$42 million in research grants in 2002 and may win further funds in 2004. Part of the military's interest concerns its troops abroad, who may have been exposed to mad cow disease. Prion diseases are also a potential bioterror agent. ■

♦ www.nap.edu

UK considers plans to shake up clinical trials

Jim Giles, London

An ambitious proposal was made this week to reinvigorate clinical trials in the United Kingdom. And the government looks likely to implement it, at least in some form, in a bid to strengthen Britain's biotechnology and pharmaceutical industries.

The plan would set up a National Clinical Trials Agency with an annual budget of £150 million (US\$250 million). The agency would fund clinical research and advise pharmaceutical companies on the conduct and regulation of clinical trials. It would also help the National Health Service (NHS), which runs most British hospitals, to find patients to take part.

The report from the Bioscience Innovation and Growth Team — made up of some 70 researchers, industry representatives and government officials — was released on 17

November. It recommends boosting funds for research and development in the NHS from 0.9% (currently £560 million) of its total budget to 1.5%. The report also suggests creating 100 places a year on degree courses that would combine medicine and research, similar to MD-PhD degrees in the United States.

The government is to respond within six months. But team members are confident: Tony Blair, the prime minister, has repeatedly pledged support for the bioscience industry.

Some of the group's plans may prove too controversial to implement, however. It suggests reducing from 20 to 3 the number of protesters required for the police to put restrictions on an animal-rights demonstration. Science minister David Sainsbury says he doesn't see an immediate need for this. ■

♦ www.bioindustry.org/bigtreport

Cornell axes Elsevier journals as prices rise

Jonathan Knight, San Francisco

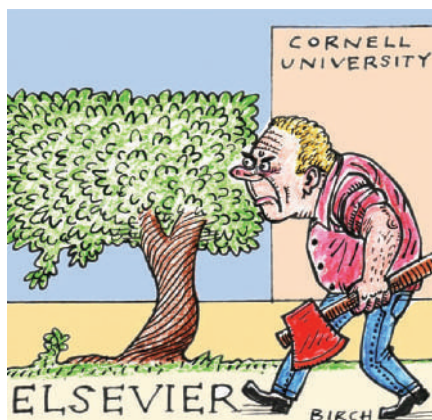
A top US research university is set to cancel its subscriptions to several hundred scientific journals published by Elsevier in January, in response to spiralling subscription costs.

The decision by Cornell University in Ithaca, New York, released in a statement last week, caught the attention of library officials at other US research universities who say that they may follow suit.

Netherlands-based Elsevier, which owns a quarter of the global market in scientific and technical journals, played down the importance of the move. Eric Merkel-Sobotta, its head of public relations, declined to comment on the company's negotiations with Cornell, but said that the firm is working hard to accommodate all of its customers. He added that Elsevier's subscription rates are rising by less than 7% annually, an increase that is necessary to cover the cost of expanding journal content.

Like other large publishers, Elsevier offers 'bundled' subscriptions to its journals along with electronic access. Many institutions have signed such agreements to gain access to large numbers of journals.

"This started out as a very good deal for universities," says Ted Bergstrom, an economist at the University of California, Santa



Barbara, who studies journal pricing. But the cost of such arrangements has risen faster than the rate of inflation, and economic woes have put library budgets under pressure.

At Cornell, the increases have forced the library to cut back on its non-bundled titles — but bundles cannot be touched without abrogating the original pricing deal.

Cornell's deal with Elsevier, now priced at \$1.7 million, consumes a fifth of the university's total periodical budget. When the library tried to cancel individual Elsevier titles, university officials say, the prices of the remaining titles increased significantly,

offsetting any savings. "To save a little, you have to cancel a lot," says Cornell's associate collections librarian, Ross Atkinson. Cornell will now return to a title-by-title plan with a vastly reduced number of journals, he says.

Cancellations by other universities are also likely, says Duane Webster, director of the Association of Research Libraries in Washington DC. "Cornell is just the first," he says.

Among those still negotiating is Harvard University, which is unlikely to renew its deal with Elsevier, according to library director Sidney Verba. He says that the price rises will probably result in a large reduction in Elsevier subscriptions.

The University of California has been in negotiations with Elsevier since March over the rising price of online access. In a notice to faculty members dated 15 October, the university's academic senate warned that a reduction in journal access would be likely if no agreement could be reached.

Last month, the faculty senate at the University of California, Santa Cruz, said that Elsevier's pricing was unsustainable. It urged faculty to give up editorships at Elsevier journals and to submit papers elsewhere. Such pressure may force Elsevier to give ground, Bergstrom predicts. "We are going to have an interesting show to watch," he says. ■

Commercial copying charges hit European academics



Double take: making copies for commercial use will now carry a fee on top of subscription costs.

Jim Giles, London

European academics are chafing under a regulation that will require some of them to pay €20 (US\$24) every time they photocopy a journal paper.

A May 2001 European Union directive on copyright, already implemented in Germany, Britain and four other nations, requires the payment if any "commercial" use of the work is anticipated. Under UK guidelines, such commercial use includes research in writing a book or preparing an article for which the author will be paid.

The UK guidelines, jointly issued in early October this year by the British Library and the London-based Copyright Licensing Agency, say that the fee could also pertain to copies made by students who are sponsored by companies.

"This will be a nightmare for researchers if it is strictly applied," says Roger Elliott, chair of the intellectual-property committee of All European Academies, an Amsterdam-based federation of science and humanities academics. The size of the fee is set by journal publishers, but

generally ranges from €7 to €40 per copy.

In Germany, where the rules came into force in September, librarians are unclear about what their implications will be. "I'm sure librarians will allow copying in the grey areas," says Harald Müller, library director at the Max Planck law institute in Heidelberg. "But they are a bit scared. They could be breaking the law."

Many library staff say it is too early to assess the impact of the change. But at the Royal Society of Chemistry in London, librarian Nigel Lees fears that the society's document-delivery service is at risk. The service supplied 600 documents a month before the change, but Lees fears that requests will drop once regular customers see their invoices. "I think it is totally unnecessary," says Lees. "We pay once for subscriptions and then the customer pays again."

Publishers say that the intent of the law is to focus on copying by commercial companies. But if academics bear the brunt of the change, publishers could face a backlash, some observers say. "This is another impediment to research by intellectual-property rights," says Elliott. "And another boost to open-access publishing." ■

Task force set up to combat threat of political interference

Erika Check, Washington

One of the largest health-research institutes in the United States has set up an informal task force to monitor political threats to its funding.

Al Sommer, dean of the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, has assembled a committee to collect information about a 'hit list' of grants funded by the US National Institutes of Health (NIH). Four members of the school's faculty were named in the list, which was given to a US congressional committee by a religious lobby group.

The move by Hopkins researchers shows how deeply the issue has affected the research community. The lobby group has urged lawmakers to withdraw funding for the grants on its list, which include research on topics such as drug use, prostitution and homosexuality.

Congress has so far not revoked funding. But this July, the legislature narrowly voted down an attempt to halt funds for a smaller set of NIH-funded projects, so scientists are taking the 'hit list' seriously.

Sommer says that the four-person task force at the school of public health will "serve as a focal point through which faculty and students can report instances about which they might be concerned", such as phone calls from federal officials or the media. Although the task force has not decided to lobby, Sommer does not rule out further action. "We're trying to collate what's happening so we can form an institutional response," he says.

Hopkins gets more NIH funding than any other establishment of its type. Last year it received \$88 million from the NIH — 44% of its research money, and about a quarter of its total budget.

So far, no other institution has taken such an organized approach, although many have been affected. The University of California, for instance, had 31 researchers on the list. But Robert Dynes, president of the University of California, has written to the state's congressional delegation on behalf of its researchers, some of whom think that this will head off the political threat.

"We have faith in the peer-review process, and that it will ultimately be viewed as an objective way of considering what is proper science," says Cynthia Gómez, co-director of the Center for AIDS Prevention Studies at the University of California, San Francisco. ■



Hans-Heinrich Trute: rules made to protect the accused leave those who speak out in danger.

Whistles blow in vain on bad practice in German research

Alison Abbott, Bonn

In summer last year, a consultant dermatologist at the University of Göttingen's skin clinic was stripped of responsibilities and demoted to a technical position. His crime? Pointing out that a patient in a neurodermatitis therapy test, which was deemed successful, had been symptomless before the treatment was given. Blowing the whistle on research misconduct can be a dangerous game in Germany.

Such incidents were on the minds of university ombudsmen from across Germany who gathered in Bonn last week to review their progress in stamping out misconduct. Many of them said that whistleblowers remain poorly protected in Germany, and that sloppy research practices, particularly in the clinical arena, are still rife.

The ombudsmen were hired as part of an effort by the DFG, the German research funding agency, that began in earnest in 1997, when a serious fraud case emerged involving two clinical researchers who fabricated or manipulated data in nearly 100 papers (see *Nature* 405, 871–872; 2000).

Back then, the DFG issued guidelines stating that universities should teach the rules of scientific good practice, appoint ombudsmen to act as independent mediators in cases of conflict and establish procedures for investigating such allegations.

By 2001 it had to threaten to withhold funding from non-complying institutions. The DFG also appointed three ombudsmen of its own, to whom scientists can turn if they do not want to discuss their concerns locally.

These catalogued 74 cases during 2001

and 2002; no one has yet tallied the number handled by university ombudsmen. Most concerned disputes over authorship, with disputes over access to data, accusations of data manipulation, and alleged blocking of an individual's career also figuring prominently. Only a couple of cases were referred to formal investigation committees.

The DFG ombudsmen's cases have been published (without naming individuals) at www.rz.uni-hamburg.de/dfg_ombud. This measure will now be extended to university ombudsmen.

But the Bonn meeting exposed many frustrations, particularly about protection of whistleblowers. "Rules were designed with protection of the accused in mind," says DFG ombudsman Hans-Heinrich Trute, a law professor at the University of Hamburg. A working group will now be set up to make recommendations on protecting whistleblowers. One possibility, says Trute, could be to require universities to track their subsequent career paths, to highlight any attempt to undermine or demote them.

"The sloppiness in clinical research is so vast that we need to be looking at very serious preventive measures," says Irmelin Probst, an ombudswoman at the University of Göttingen, which has been hit by more clinical-research scandals than any other German institute. "The extreme hierarchy in old-fashioned clinical departments, and ignorance about research methods among older clinical scientists, is to blame," she adds.

Ombudsmen at the meeting pledged to persuade more universities to teach good scientific practice. ■

Whale of a catch blows hole in family tree

David Cyranoski, Tokyo

A whale carcass swarming with maggots has made a splash in the world of marine-mammal research. The rotting remains were found by marine biologist Tadasu Yamada on a trip to a remote island in the Japan Sea in 1998 to examine a whale that died in a collision with a fishing boat.

Five years on, Yamada, of Tokyo's National Science Museum, and colleagues have shown that the whale represents a new species. The finding helps to resolve long-standing confusion about how many whale species there are. But it could create problems for whale conservation.

Scientists have long suspected that there are more than just the six recognized species of baleen whales of the genus *Balaenoptera*. Researchers thought that eight whales taken by Japanese research whaling boats in the late 1970s, for example, were part of a unique species — but they couldn't prove it.

Shiro Wada, a molecular biologist at Japan's National Research Institute of Fisheries Science in Yokohama, analysed enzymes in the liver and muscle of those specimens and found dramatic differences from common Bryde's whales (*Balaenoptera brydei*) which they resembled.

"I was sure that they were a different species, but that was a period when people didn't trust comparative enzyme studies for evolutionary studies," says Wada. The full carcasses of the whales were not kept for further study.

So when the 1998 whale carcass turned up, Wada, Yamada and their colleagues com-



Analysis of the skull of a baleen whale found in 1998 suggests that it represents a new species.

pared the structure of the skull and baleen plates — sieve-like structures that baleen whales use to sift food — along with DNA from all nine whales, with other *Balaenoptera* whales. They conclude that all nine are members of a previously unknown species. They also confirm that Eden's whale (*B. edeni*) is distinct from Bryde's whale, bringing the total number of *Balaenoptera* species to eight.

The identification of the new species, announced on page 278 of this issue, should help inform estimates of whale abundances, says Justin Cooke, senior scientist at the Centre for Ecosystem Management Studies near Freiburg, Germany, and a member of the World Conservation Union's Species Survival Committee. The union does not know how many Bryde's whales there are, because of confusion over the number of species.

The Japanese whaling community, which catches about 50 Bryde's whales in the west-

ern northern Pacific each year for research purposes, estimates that there are 22,000 Bryde's whales in that part of the ocean.

Researchers at the Institute of Cetacean Research in Tokyo, who carry out much of the research whaling, deny that they are catching the new species. "We only catch the common Bryde's whale," says Luis Pastene, a population geneticist at the institute.

But Scott Baker, who studies whale population genetics at the University of Auckland, New Zealand, reckons that they can't be so sure. "Japan has been conveniently ignoring the taxonomic uncertainty," says Baker. "Until you understand the taxonomy, you can't be sure how many are out there or what you're hunting."

Baker adds that genetic methods are likely to uncover more 'hidden' biodiversity in the oceans over the next few years. But he doubts that there are many more baleen species left to find. ■

T. YAMADA, Y. TAJIMA & K. ARAI

Europe urged to counter nuclear proliferation threat

Geoff Brumfiel, Washington

Nuclear reactors used for research in eastern Europe pose a significant proliferation risk, which the European Union (EU) should take much firmer action to diffuse. That is the message a founder of US non-proliferation programmes will deliver to a meeting of European politicians in Strasbourg this week.

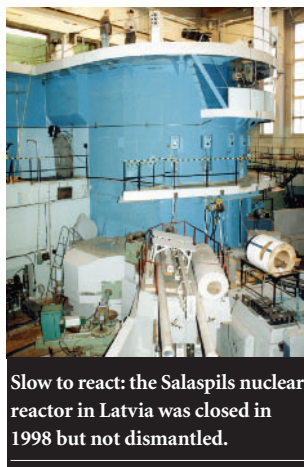
Sam Nunn, a former senator from Georgia who co-authored a 1991 bill that committed billions of US dollars to securing the former Soviet Union's nuclear-weapons stockpile, says that Europe should take the lead in dismantling or safeguarding the dozen or more Soviet-built research reactors.

The reactors are fuelled by highly enriched uranium, which could be used in nuclear weapons. "It tends to be stuff you could carry out in a backpack or pickup truck," says Matthew Bunn, a proliferation expert at Harvard University's John F.

Kennedy School of Government in Cambridge, Massachusetts.

Nunn is attending an interparliamentary conference on non-proliferation in Strasbourg on 21 November, and says that he hopes to get the EU to address the issue more forcefully. "This issue should be of acute interest to our EU allies and friends," he says.

But Bruno Tertrais of the Foundation for Strategic Research in Paris, points out that non-proliferation issues tend to be seen in Europe as a matter for the United States and Russia to resolve. "It's extremely difficult to make our opinion and political leaders understand that this is a



Slow to react: the Salaspils nuclear reactor in Latvia was closed in 1998 but not dismantled.

European issue," he says.

The low visibility of the problem in Europe has led to scant funding to counter weapons of mass destruction. Bunn says that between 1992 and 2000, Europe spent less than US\$500 million on non-proliferation, whereas the United States committed more than \$6.5 billion.

Research reactors would be a good focus for EU anti-proliferation efforts, says Michèle Flournoy of the Center for Strategic and International Studies in Washington DC. Several of the EU's new members in eastern Europe have reactors that house highly enriched uranium, says Flournoy. ■

V. LISTISIN/TITAN-TASS

Two teams make quantum leap in quest for merged molecules

London Two teams of physicists have announced the creation of an elusive form of matter called a molecular Bose–Einstein condensate.

A Bose–Einstein condensate is a group of identical particles that occupy the same quantum level and so behave as if they were a single particle. Condensates made from atoms were first created in 1995, but repeating the trick with molecules has proved more difficult.

“This is going to cause a lot of excitement,” says physicist Randall Hulet of Rice University in Houston, Texas. “We’ve all been racing to make a condensate from molecules.”

Rudolf Grimm of the University of Innsbruck, Austria, and his colleagues used lasers to cool a gas cloud of lithium atoms, as they report in a *Science* paper published online last week (S. Jochim *et al. Science* 10.1126/science.1093280; 2003). As the temperature fell, the atoms formed weak bonds, pairing up into lithium molecules and forming a condensate that lasted 20 seconds.

Meanwhile, Deborah Jin of the University

of Colorado in Boulder and her colleagues created a condensate from potassium molecules. They will publish their results in an upcoming issue of *Nature*.

Islands make waves for Chinese space base

Tokyo The tiny republic of Kiribati in the Pacific Ocean looks set to disrupt China’s space programme.

Earlier this month, the republic decided to establish diplomatic relations with Taiwan. Although the islands that make up Kiribati have a population of less than 100,000, and together cover just 717 square



kilometres, the decision could force China to close down one of the bases for its space programme.

China established a base on Kiribati in 1997 to help it observe satellites. Last month, the base was one of many used to watch China’s first manned spaceflight. It is rumoured that the base is also used by China to spy on an American ballistic-missile base in the nearby Marshall Islands.

China usually refuses to keep relations with countries that have formal ties with Taiwan and is expected to cut ties with Kiribati if it doesn’t change its mind. China has told Kiribati to reverse its decision or “face serious consequences”, according to *Xinhua* news, the Chinese government’s official English-language newspaper.

Satellite mission is star buy in online auction

Washington Online auction addicts tired of the usual Star Wars memorabilia and vintage video games got to bid on something more exciting this week — a satellite mission.

The company SpaceDev, based in Poway, California, posted a complete mini-satellite mission — featuring up to a year of orbiting time — on eBay on 10 November, with an asking price of \$9.5 million. The six-year-

old company usually survives on small contracts from the US Air Force and others. One such project saw it team up with the University of California, Berkeley, to build CHIPSat, which was launched by NASA last January. Even so, the company posted a \$208,000 loss last quarter.

Six days into the ten-day auction period, only one person had posted a bid: venture capitalist Michael Potter of California-based company Paradigm Ventures, who offered just \$250,000. He says he will use the satellite as either a broadband relay or an Earth-sensing device if he wins.

SpaceDev has not disclosed its 'reserve' price — the bid below which it will not sell — but the auction site notes that the current leading bid does not meet this sum. For comparison, bidders may note that CHIPSat cost NASA \$6.8 million.

'Money down the drain' fears for AIDS vaccine trials

Washington An AIDS vaccine candidate has failed its second large clinical trial, causing concern that a third trial now in progress could prove a waste of money.

The California-based company VaxGen announced on 12 November that its AIDS vaccine candidate — AIDSVAX — failed to protect 2,546 drug users in



Plunging fortunes: AIDSVAX's poor results have led to doubts over continuing clinical trials.

Thailand from HIV, the virus that causes AIDS. Earlier this year, the vaccine failed a similar trial in North America (see *Nature* 421, 877; 2003).

But the US National Institute of Allergy and Infectious Diseases, the US Department of Defense and the Thai government have already started another clinical trial with AIDSVAX, this time in combination with another vaccine candidate — Aventis Pasteur's ALVAC. The trial will involve 16,000 volunteers and cost \$120 million.

The trial's sponsors say there is evidence that the combination could prove better than either vaccine alone. But John Moore of Cornell University's

Weill Medical College in New York says this evidence is "extremely poor". A lot of people are concerned that the money will be washed down the drain, he adds.

Cancer grant chief denies Harvard bias

San Francisco A Congressional committee is investigating whether a former director of the National Cancer Institute with links to Harvard University had a hand in giving the university a \$40-million grant.

Richard Klausner, who was appointed to the institute by the Clinton administration, was director when the grant was being awarded — it eventually went to Stuart Schreiber's genetics lab at Harvard in 2002. After Klausner left the institute in 2001, he became a paid consultant to Schreiber's firm, Infinity Pharmaceuticals.

The House Committee on Energy and Commerce, led by Representative Billy Tauzin (Republican, Louisiana), has asked Harvard University and the National Institutes of Health to provide records relating to the grant.

Klausner says he had no say in the grant's destination. He excused himself in 1999 from any decisions relating to Harvard, as he was standing as a candidate for president of the university.

It's a scoop!

In the highly competitive world of cell and molecular biology, there are no prizes for coming second. But is the pressure to be the first to publish 'hot' results distorting scientific progress? Helen Pearson investigates.

BILL OSMENT

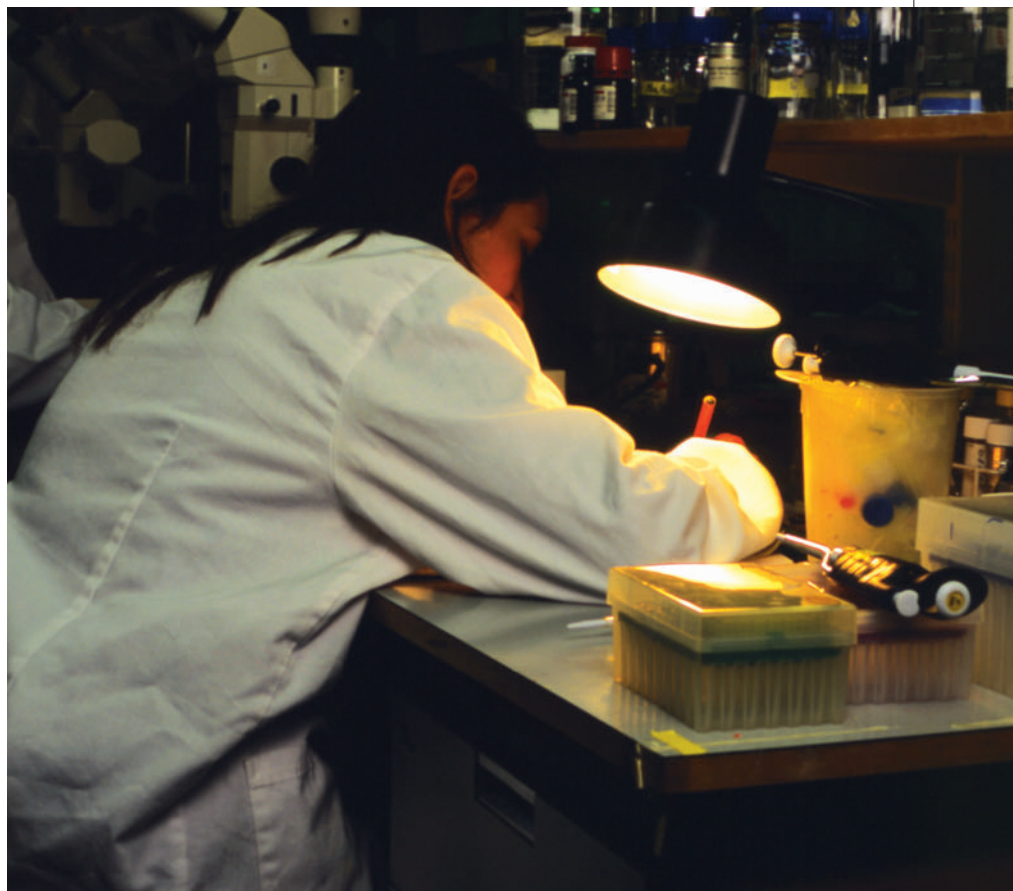
In 1998, Robert Insall abandoned his own lab and set up camp in that of his wife, cell biologist Laura Machesky. For three months, the pair slaved into the night at University College London. Their goal was to confirm a tantalizing link between biochemical signals entering a cell and the proteins that direct the cell's movement and division. The reason for the researchers' urgency: the fear that one of two competing labs would stumble on the same findings and 'scoop' them.

In the end, the effort paid off. Competitors cursed as Machesky revealed her findings at that year's meeting of the American Society for Cell Biology (ASCB) in San Francisco. At the poster summarizing her talk, someone whipped out a mobile phone and started reading the methods to researchers in their lab. Machesky and Insall were the first to publish their findings¹, but within months, four other papers had appeared describing similar results²⁻⁵.

Insall and Machesky's determination to be first may seem extreme, but such obsession is not unusual among cell and molecular biologists. The stakes are high: peer recognition, a grant — even a sought-after faculty position — can be won or lost on the back of a ground-breaking paper. And unlike some other fields, where gathering data takes years, it is sometimes possible to turn round an experiment, get it written up and submitted to a journal in a matter of days or weeks. With today's off-the-shelf laboratory kits, plus instant access to genome data, many experiments can rapidly be copied. So it's not surprising that biologists following a hot lead tend to play their cards close to their chests.

Competitive streak

The race to publish first has always been there, and is generally thought to help accelerate scientific progress. But many cell and molecular biologists are worried that things are reaching such a frenzied pitch that cracks are beginning to show. At some scientific meetings, for example, mutual suspicion is drying up the free exchange of ideas. More pressingly, some researchers fear that spiralling competition is encouraging the rapid publication of 'quick and dirty' experiments — a trend that is now being abetted by the swift turnaround of papers facilitated by electronic submission and online publication. "I don't think there's time to do things carefully," says Steven Reed, who



works on the cell cycle at the Scripps Research Institute in La Jolla, California.

Ambitious scientists may drive much of the competition, but researchers say that journals must also shoulder some of the blame. They argue that top journals fuel competition by milking the research community for novel results — and, according to some biologists, are sometimes prepared to let standards drop in order to publish a particularly hot finding. "There's more emphasis on novelty than if you've really done a good job," complains Stephen Schoenberger at the La Jolla Institute for Allergy and Immunology.

Competition is nothing new in science, but the available data seem to support the impression that it is becoming more intense. One survey⁶ found that the number of scientists who were unwilling to talk openly about their work rose from 50% in 1966 to 74% in 1998. In addition, more than three-quarters of biologists surveyed in 1998 expressed the fear that they might be scooped. "There's a trend towards increasing paranoia," agrees

Tyler Jacks, a cancer researcher at the Massachusetts Institute of Technology.

Perhaps the most obvious casualty is the 'buzz' surrounding new results presented at scientific meetings. Many researchers tell tales of revealing results in a PowerPoint presentation, only to see similar findings materialize in a competitor's paper just weeks later. "It is absolutely routine to be squeezed for information at a meeting by someone, only to find out that they are working on the same thing without revealing it," says Karel Svoboda, a neuroscientist at Cold Spring Harbor Laboratory on Long Island, New York.

Scared by the threat of a scoop, some tight-lipped biologists refuse to present work that is not already safely published or in the press. Novel results are getting so thin on the ground that some frustrated researchers are abandoning major talks at big gatherings, such as the annual meeting of the ASCB, preferring to share ideas at smaller meetings, in the bar, or not at all. "It's really terrible. Competition has almost made conferences obsolete," laments Thomas Maniatis, a



Burning the midnight oil: some molecular and cell biologists now feel compelled to work unsociable hours to get the data needed to put them ahead of the competition.

molecular biologist at Harvard University.

While some researchers fret about competition stifling communication at meetings, its impact on published research is also giving cause for concern. Some biologists say that the constant battle for the scoop means that they are being forced to cut their experiments short in order to publish before their rivals—a trend that they fear is starting to undermine the quality of the scientific literature.

Race against time

Cases in which researchers and journals have raced each other are not hard to find. In January this year, for example, a team led by Francis Collins, director of the National Human Genome Research Institute in Bethesda, Maryland, submitted a paper to *Nature* that pinpointed the gene underlying a rare disorder called Hutchinson–Gilford progeria syndrome, which causes sufferers to age prematurely. The gene, called *lamin A*, encodes proteins that are building blocks of the cell nucleus.

Collins says that publication of his study

was delayed because a second manuscript arrived at *Nature* in February reporting a mouse model of progeria⁷. *Nature's* usual policy is to publish such closely related papers together. But then Collins heard of a paper scheduled for publication in *Science*, identifying mutations in the *lamin A* gene in two progeria patients. “That was a stunner,” he says.

The lead researcher on the *Science* paper, Nicolas Lévy of INSERM, the French medical research agency, in Marseille, says he had suspected that there was an overlapping paper in the press. Lévy says that this led him to write his paper in a matter of days, but he is adamant that he didn’t rush his experiments.

Anxious to avoid being scooped, Collins called Lévy to ask if he would be willing to announce the results simultaneously—and asked *Nature* to accelerate his own study’s publication. *Science* received Lévy’s paper⁸ on 4 March, accepted it on 1 April and published it online on 17 April. Collins’s *Nature* paper⁹ appeared online eight days later. But the journal bent its normal rules to allow Collins to announce his results on the same day that Lévy’s paper came out.

Other researchers argue that such competitive antics are far from ideal—and in some cases might cause incomplete or sloppy work to be published. “It’s not a great way to do the best science,” says Susan Michaelis, a cell biologist at Johns Hopkins University in Baltimore, Maryland.

The electronic age has lent such contests

a new sense of urgency. Where once authors considered themselves relatively safe in the weeks or months between submission and publication, the advent of electronic submission, e-mail reviews and online publication mean that a competing paper can be reviewed, revised and published in a matter of days. “It ratchets things up a bit,” says Stuart Firestein, who studies smell receptors at Columbia University in New York.

Whether increasing competition and rapid online publication are conspiring to undermine standards remains a matter of opinion. But some biologists admit that there is always the temptation to cut a few corners to get work published quickly. “It’s a gamble,” says Larry Katz, a neuroscientist at Duke University in Durham, North Carolina. “Do I make it a great paper or do I get it out the door?”

Exclusive club

Many researchers believe that leading journals such as *Nature*, *Science* and *Cell* could help to ease excessive competition by relaxing their demands for exclusive results. At *Nature* and *Science*, for instance, a paper will generally be turned down if it arrives a few weeks after one with largely overlapping conclusions has come out, even if it is of comparable quality. Some biologists argue that this is too harsh—a second or third paper may actually be more thorough and influential to a field. “It should be possible to publish work that isn’t neck and neck,” says Reed.

Biology editors at *Science* and *Nature* claim that a strict criterion of novelty is necessary, given the enormous number of manuscripts that the journals receive. *Nature's* chief biological-sciences editor, Ritu Dhand, denies that the journal sacrifices quality in order to get a scoop. “We would not skip a re-review or revision if it was deemed critical to the conclusion,” agrees *Science's* deputy editor for life sciences, Katrina Kelner. Editors at *Cell* declined to comment.

But ultimately, some biologists argue that it is up to individual researchers not to be drawn into a cycle of excessive competition. Really sloppy work will ultimately lose you your collaborators, respect and reputation—the very things you are competing to earn. “A difference in days or weeks doesn’t matter,” says Keith Yamamoto, a molecular biologist at the University of California, San Francisco. “It’s how good the work is.” ■

Helen Pearson is a reporter with *Nature's* online news team, based in New York.

1. Machesky, L. M. & Insall, R. H. *Curr. Biol.* **8**, 1347–1356 (1998).
2. Madania, A. *et al. Mol. Biol. Cell* **10**, 3521–3538 (1999).
3. Winter, D., Lechler, T. & Li, R. *Curr. Biol.* **9**, 501–504 (1999).
4. Yasar, D., To, W., Abo, A. & Welch, M. D. *Curr. Biol.* **9**, 555–558 (1999).
5. Rohatgi, R. *et al. Cell* **97**, 221–231 (1999).
6. Walsh, J. P. & Hong, W. *Nature* **422**, 801–802 (2003).
7. Mounkes, L. C., Kozlov, S., Hernandez, L., Sullivan, T. & Stewart, C. L. *Nature* **423**, 298–301 (2003).
8. De Sandre-Giovannoli, A. *et al. Science* **300**, 2055 (2003).
9. Eriksson, M. *et al. Nature* **423**, 293–298 (2003).

A continent divided

African activists, backed by wealthy supporters in the United States and Europe, are locked in combat over the merits of transgenic crops. Ehsan Masood tracks the people, the politics and the cash behind the campaigns.

B. PRESS/PANOS PICTURES



Appealing to the people: Africa's food shortages have put it at the heart of a political tug-of-war over genetically modified crops.

When Jocelyn Webster was asked by a South African reporter for her opinion on “genetically modified orgasms”, she was exasperated but not surprised. For Webster, who heads the pro-biotech campaigning group AfricaBio, this question was just one more symptom of the endemic misunderstanding about transgenic technologies in Africa.

Africa is emerging as one of the front lines in the battle for acceptance of genetically modified (GM) foods. Webster believes that transgenic agriculture is vital in the fight against world hunger, and AfricaBio, along with agribiotech companies and other pro-biotech campaigners, is now fighting tooth and nail, often by somewhat controversial methods, to spread the word about GM crops. But the anti-GM lobby is equally powerful and vociferous, and vast amounts of money are flowing in to Africa in support of both sides of the argument, as the various

parties try to influence policy-makers and the public.

For AfricaBio, a coalition of scientists and companies based in South Africa, the idea is to improve GM's image — perhaps with good reason. Today, some 200 million people in sub-Saharan Africa don't know where their next meal will come from, and the problem will not be going away. Despite international aid to feed the hungry, Africa will still have 183 million undernourished citizens by 2030, according to a report published by the UN Millennium Hunger Task Force this year. AfricaBio is one group among many that believes transgenic crops, modified so that they will grow in salty soils or conditions of drought, offer a solution to starvation.

But the group's methods would be considered in some countries to be blatant

media manipulation. Webster talks about training journalists how to report GM stories, telling them that the term ‘genetically improved’ is more accurate than ‘genetically modified’. In early 2003 she hosted a press

“We resent the way that the stereotyped image of the hungry has been used to force a style of agriculture that will only exacerbate problems.”

Tewolde

briefing where the journalists were fed GM fritters without knowing it. The idea, Webster says matter-of-factly, was to demonstrate that GM food tastes just the same as conventional fare, and does no harm. She claims that the journalists were amused and there were no

angry headlines in the next day's papers.

Although Webster stresses the role of GM crops in improving nutrition in Africa, there are wider issues at stake for companies such as the US-based agribiotech giant Monsanto, which is one of the funding sources of AfricaBio. If GM crops can be sold as the way to feed the starving, there could be a subtle

shift in the political landscape worldwide, making GM food more acceptable to consumers in Europe and elsewhere.

On the other side of the fence, opponents of GM crops are just as determined to keep them out of Africa. Over the past few years, international non-governmental organizations (NGOs) involved in development, such as Oxfam, Christian Aid and Action Aid, have joined their environmentalist cousins from Greenpeace and Friends of the Earth in lobbying against GM in both Europe and Africa. Their extensive network of international media contacts has helped them to generate publicity for their views in a relatively short time.

Battle lines

The two sides are squaring up to each other at a crucial time for Africa. South Africa is the only country on the continent that is currently growing GM crops commercially. Egypt, Zimbabwe, Kenya and Uganda are doing GM research. But elsewhere, GM is still very much an emerging issue. Only three countries so far — Egypt, South Africa and Zimbabwe — have introduced regulations to govern transgenic agriculture.

The local policy-makers who will ultimately decide on the future of GM in Africa are being pushed and pulled in both directions, and are being showered with money from the developed world, some in the form of grants for 'biosafety research' — money for regulatory infrastructure and testing of whether GM affects the environment or human health.

At last year's Earth Summit in South Africa, the US government pledged \$100 million over ten years to support agricultural biotechnology in the developing world. In May, the US Agency for International Development (USAID) announced a grant of \$15 million to support biosafety policy-making and research in Asia and in East and West Africa. At the beginning of October this year, the Bill & Melinda Gates Foundation — the charity established by Microsoft founder Bill Gates — pledged \$4 million for GM technologies, as part of a \$25-million project to combat malnutrition. And later the same month, Germany approved a grant of €2 million (US\$2.3 million) to help African countries strengthen biosafety laws.

The USAID and German grants are particularly significant when viewed in the context of a looming trade dispute between the United States and the European Union (EU) over the latter's failure to approve new GM products for sale and growth (see *Nature* 425, 655; 2003). Against this background, both powers are trying hard to influence African interpretations of a new international treaty on GM trade — the Cartagena biosafety protocol, which came into effect on 11 September 2003. The protocol governs trade in 'living modified



Opposite sides: Florence Wambugu (below) is a firm supporter of GM technology, whereas Tewolde (above) remains sceptical of its benefits.



organisms' — from seeds to fish — intended for direct release into the environment. A country can use it to block GM imports if it thinks a crop will damage the environment, human health or the trade in locally produced goods.

Vested interests

The United States and its private-sector allies would like the laws in African countries to reflect their own views — that GM technology is inherently safe unless proven otherwise, and that countries should not be allowed to refuse GM imports just because they don't particularly want to eat GM food. There are strong reasons to think that the USAID grant will be used to support this position. In April, the agency's head, Andrew Natsios, told the US Congress: "The great bulk of African agricultural ministers, presidents and prime ministers I have spoken with are all interested in bringing this technology into their agricultural systems." Natsios accused "Europeans" of spreading

misinformation, which did "enormous damage to poor people in Africa".

European governments and aid groups see things differently. The EU, unlike the United States, insists that all GM produce is labelled and traceable back to its source. Measures that would add those requirements, and even stricter regulations, to the Cartagena protocol were endorsed by the African Union's heads of state at a meeting in Mozambique this August. The German grant will help member states to draw up similar regulations that would allow them to block imports of all GM produce without needing to cite specific reasons, and possibly to claim damages if a GM product harms human health or the environment.

The brains behind this African law, and the man who is courting European funding for the project, is Tewolde Berhan Gebre Egziabher, head of Ethiopia's Environmental Protection Authority. Tewolde is soft-spoken and understated; it is hard to believe that his influence extends across Africa and beyond. He has powerful European supporters in environment ministries and NGOs, including former British environment minister Michael Meacher, who since his departure from government has emerged as a figurehead for the anti-GM movement. Meacher hosted Tewolde at the House of Commons in London last month to address members of parliament on the African position regarding Cartagena.

Natural response

When *Nature* caught up with Tewolde at his London base, the offices of the environmental organization the Gaia Foundation, he argued that he isn't entirely anti-GM, nor does he want to stand in the way of the technology feeding the poor. "We simply do not want to grow GM crops without due consideration given to human health, domestic animals and the environment," he said over a plate of non-GM rice, bread and lentils. "We suspect that Africa is high on the agenda for the United States' next push for GM acceptance. And we resent the way that the stereotyped image of the hungry in developing countries has been used to force a style of agriculture that will only exacerbate problems of hunger and poverty."

Until the late 1990s, those views dominated nascent African public opinion on transgenic agriculture. Environmental NGOs drip-fed messages questioning the safety of GM technology, its relevance to the needs of ordinary Africans, and the intentions of the multinational corporations who promote it. Their argument was, and is, simple: Africa does not need to waste its scarce resources on a technology whose principal products to date are herbicide-tolerant cotton and longer-lasting tomatoes. The anti-GM campaign reached a high point in 2000, when European and African

governments, led in part by Tewelde, formed an alliance to defeat the United States and other grain-exporting countries at the United Nations and forced through the Cartagena protocol.

Stung by this defeat, pro-GM campaigners realized that, if they wanted to wield equal influence in Africa, they would need to take a leaf out of the green activists' handbook and bring on board successful public-relation strategies. The biotech lobby has since sought out charismatic African scientists, farmers and policy-makers who believe in the power of GM, and has built up their public profiles at home and abroad.

A prime example is Kenyan plant scientist and farmer's daughter Florence Wambugu, who in 2001 formed A Harvest Biotech Foundation to spread the message that Africa wants GM crops. She has strong connections to Monsanto, having done her postdoctoral research at its labs in St Louis, Missouri.

A boost for biotech

Wambugu, a passionate and assertive campaigner, wrote to a committee of the US House of Representatives about GM and poverty earlier this year. She argued that the European anti-GM lobby's primary accomplishment was to "keep safe and nutritious food out of the hands of starving people". Those are strong words, but her choicest remarks are directed at Tewelde. "This is a very exciting time for biotechnology in Africa, but Tewelde's agenda is to prevent Africa's participation in this. My agenda is to promote it," she says.

AfricaBio is singing from the same hymn sheet. When this article was being researched, Webster was visiting Germany and Britain, where Tewelde's influence is strong, trying to counteract the "nonsense", as she puts it, that he and his allies promote. Webster was accompanied by Thembitshe Joseph Buthelezi, a cotton farmer and chair of a federation of small farmers' associations in South Africa's eastern KwaZulu-Natal province. Over a breakfast meeting in London organized by Monsanto, the South African pair enthused about the power of GM to reduce poverty.

Buthelezi argues that planting *Bt* cotton — a GM variety that produces its own insecticidal toxin — has transformed his life. He says that, in common with 90% of the region's cotton farmers, he is seeing better yields, a better quality of cotton, and is spending less on the chemicals and labour needed to spray insecticides. One of the slides that he uses in his presentations to governments and corporations proudly says: "Normally, I used to ask my wife how we intend to pay our bills. Now I ask her,



End of the line: could GM crops stop the queues for food in Africa?

how are we going to spend this money?"

Taking such feel-good stories to consumers and the media in Africa and abroad is an important plank in AfricaBio's strategy. To that end, it is helping to train staff working in South Africa's supermarkets — including the UK-based Tesco chain — to handle questions about GM foods from shoppers. The organization is also working with women's groups in poor townships, and is advising the government of Lesotho — a tiny independent country landlocked within South Africa — with its planned biosafety legislation.

Perhaps surprisingly, considering AfricaBio's stance, not all of its funding comes from sources with a strong pro-GM agenda. One of its biggest donors is the Rockefeller Foundation — the wealthy US-based philanthropic organization set up by John D. Rockefeller, founder of Standard Oil. The Rockefeller Foundation has positioned itself as a bridge between opposing sides, and seems to have the trust of all the key players. It funds AfricaBio and Florence Wambugu's A Harvest. But it also gives grants to NGOs with a cautionary stance on GM, including the Winnipeg-based ETC Group in Canada and the London-based Consumers International, a federation of consumer groups and agencies worldwide.

Balancing act

Gary Toenniessen, the Rockefeller Foundation's head of food security, says it supports both sides in the GM debate to ensure that the public has access to a spectrum of information, both for and against. The hope is also to promote dialogue between opponents and help them to see an issue from another perspective. "Often they find they have more in common than is evident from their public statements," he says.

Toenniessen does not think that GM technology by itself is the answer to hunger in Africa — it is, at best, one component in an overall strategy to achieve a secure, sustain-

able food supply. This is evident in the foundation's portfolio of food-security research: it invests about \$3 million annually in genetic modification, and up to \$7 million in supporting conventional plant breeding.

The Rockefeller Foundation's approach is in tune with the thinking of an international task force assembled by United Nations Secretary-General Kofi Annan in 2000 to find ways of halving global hunger before 2015. The Millennium Hunger Task Force, overseen by Jeffrey Sachs, director of the Earth Institute at Columbia University in New York, published its

first report this April. The task force agrees that biotechnology has opened up new opportunities that could help feed Africa's starving, including projects involving drought-tolerant maize and disease-tolerant cooking bananas. But it points out that GM won't remove many of the present barriers to feeding the poor.

Between 1980 and 1995, sub-Saharan Africa was the only region in the developing world that showed a decrease in crop production. Yields increased by 27% in Asia and

12% in Latin America, but fell by 8% in Africa. The task force concludes that this is principally because of poor-quality soils, inadequate irrigation, fertilizers that are sold in remote areas at inflated prices, pot-holed roads that delay the sale of fresh produce, and little access to the

credit that would help farmers manage their businesses better. Unless these conditions are improved, the task force concludes, the GM revolution won't be able to live up to its promise.

Pro-GM campaigners have a hard time disputing these findings. Wambugu herself was a member of the task force and doesn't disagree with the facts — although she observes that the emphasis on improving Africa's soils may have something to do with the fact that the task force was co-chaired by a soil scientist. Webster also agrees that GM alone is not a panacea. But if anti-GM campaigners succeed in the battle now going on in Africa, she claims, they will stock up problems for the future. "You can say 'no' to the technology now," she argues. "But when you'll really need it, the technology won't be ready."

It is difficult to predict where Africa's GM debate will go from here. Almost certainly it will involve a legal minefield of international treaties and arguments about the economics of trade. But you can count on Wambugu, Webster and Tewelde remaining at the heart of the battle to win the hearts and minds of Africa's people.

Ehsan Masood is a freelance writer based in London.

Precautionary principle cuts costs as well as risks

Even allowing for false alarms, erring on the side of caution makes good financial sense.

Sir— In his Book Review of *Politicizing Science*, “Science exiled” (*Nature* **425**, 663–664; 2003), Paul M. Grant repeats the views expressed by several of the book’s contributors that science is often subverted by politics and that the precautionary principle is widely overused. Such an uncritical reading is not helpful to *Nature* readers.

Contrary to the alarms raised in the review about “political interference”, none of the book’s authors have been cowed by their encounters with politics. All remain outspoken, and it is hard to see even Fred Singer as marginalized when the president of the United States has rejected Kyoto and

forsworn any action to slow carbon dioxide emissions in the near future.

It is also strange that Grant points to quantitative risk analysis as a necessary antidote to overuse of the precautionary principle. In environmental regulatory practice, cost–benefit analyses rarely account for the cost of waiting for better data before acting. A recent economic study argues that a more liberal application of the precautionary principle in environmental regulation would pay for itself despite the greater rate of false alarms (S. W. Pacala *et al.* *Science* **301**, 1187–1188; 2003). The United States Office of Management and Budget has also

recently reported that the benefits of the country’s environmental regulations exceed their cost by a factor of six.

The frustrations of making public policy in a democracy have been with us since ancient Athens and will not disappear soon. In my view, a warmer embrace of both politics and the precautionary principle would do a world of good for scientists who presume to instruct the public and its representatives.

Jonathan M. Gilligan

*Department of Geology,
Vanderbilt University,
VU Station B Box 351805, Nashville,
Tennessee 37235-1805, USA*

Middle East: university funding for Palestinians

Sir— Your recent News Feature described how Israeli and Palestinian scientists link “Across the great divide” (*Nature* **425**, 444–449; 2003). I work at the Hebrew University of Jerusalem, in the department of physical chemistry, and I have been surprised by the increasing number of Palestinian graduate students (and applicants) in our department. One such student, for example, is about to conclude his postdoctoral work in experimental physical chemistry, and has recently accepted a position at Al-Quds University of East Jerusalem, where he is constructing a modern nano-chemistry laboratory.

These promising individuals are struggling not only with the impossible political situation of the Middle East, but also, and predominantly, with the hardships of everyday economic conditions. With the deep budget cuts imposed by the Israeli government, the Hebrew University has had to cut 20% of its support to all graduate students. While all our students were adversely affected, our Palestinian students are among the hardest hit. Owing to language barriers, they usually cannot assume teaching positions within our university, nor do they have affluent families to carry them through. One Palestinian student of quantum chemistry had to postpone his graduation in order to make a living through an extramural teaching position.

The Hebrew University is currently exploring several practical routes for enlisting potential contributors to assist these students. It is remarkable that the academic community which suffered most from the recent wave of terror (as reported

in your News Feature) responds in this noble manner. One hopes that when its Palestinian students graduate and assume leadership positions in Palestinian academic institutions, they too will contribute towards bridging “the great divide”.

Noam Agmon

*Department of Physical Chemistry,
The Hebrew University, Jerusalem 91904, Israel*

Middle East: nothing to fear except terrorism

Sir— With a colleague, I supervise a group of five graduate students at the Hebrew University’s Department of Pharmacology on the Hadassah Hospital campus. One of this group is an Israeli Arab. In fact, some 20% of Israel’s population is Arab, although this was not mentioned in your News Feature (*Nature* **425**, 444–449; 2003) about the status of Israel–Palestinian science cooperation and related issues.

If you visit our campus you will hear Arabic spoken and see many Arab students at all stages, including many doctoral students, some of them Palestinians from East Jerusalem and elsewhere. They come and go without any problem; no one bothers them at all. The only thing they have to fear is being blown up on a bus or in a cafe by a Palestinian suicide terrorist. The Hebrew University has a proud reputation of academic openness with regard to the Arab minority. Any boycott of Israeli academic institutions is likely to hurt these Israeli Arabs more than the Jews.

As someone who worked between 1994 and 1996 towards Israeli–Arab cooperation in biotechnology as a US member of a Trilateral (United States–Jordan–Israel)

Science Group, I cannot even visit a Palestinian university campus because of the extreme danger. In my view, until the Palestinian culture of violence changes, all attempts at scientific cooperation are illusory. Instead of supporting one-sided boycotts against Israeli academics and institutions, which are in any case in favour of peaceful solutions, any well-meaning academics who would like to see peace should be demonstrating against the terrorist groups.

Jack S. Cohen

*Pharmacology Department, Faculty of Medicine,
The Hebrew University, Jerusalem 91120, Israel*

Casting light on mystery of auroral flashes

Sir— There may be a simple explanation for the puzzling auroral flashes seen by astronaut Ed Lu (*Nature* **425**, 888; 2003). They could be due to an electrical discharge between regions of the atmosphere containing different electrostatic charges.

The fact that the flashes were visible only in the direction of the aurora suggests that the charged particles in the aurora are important. It is well known that ionizing radiation can trigger an avalanche breakdown of an insulating gas; this is, in fact, one of the ways of detecting ionizing radiation, as in a Geiger counter. The resulting burst of energy will release the built-up charge, and create a flash of light in the process. Some explanation for the initial build-up of charge is also needed, and could arise from the frictional forces of the high-speed winds in the upper atmosphere.

Len Freeman

23 Hope Street, Cambridge CB1 3NA, UK

Too young for gardening

Jim Watson's former colleagues ponder his life after the Nobel prize.

Inspiring Science: Jim Watson and the Age of DNA

edited by John Inglis, Joseph Sambrook & Jan Witkowski
Cold Spring Harbor Laboratory Press: 2003.
503 pp. \$35, £19.99

Robert Olby

When you've reached your peak, some say, it's downhill all the way. So what happens when you've won a Nobel prize? You meet the press, receive congratulations, go on the lecture circuit, and then settle for a quiet life in the lab, with time for golf or gardening and perhaps an autobiography at 65. But James Watson won his Nobel at 34 — too young for an easy-going life of gardening and golf. What better activity for such a young Nobel laureate than to promote the science he believed in so passionately — molecular biology.

First came the classic textbook (*Molecular Biology of the Gene* in 1965), then his notorious best-selling autobiography (*The Double Helix*, 1968), followed by his entrepreneurial and administrative achievement in creating that heaven-on-earth for molecular biologists at the fading institute by the sea, the Cold Spring Harbor Laboratory. Finally he had a short-lived reign over the then National Center for Human Genome Research. That's more than enough for one lifetime.

Recollections of this varied career — and in particular Watson's direction of the Cold Spring Harbor Laboratory — form the meat of this handsomely produced book. The numerous contributors show how Watson has been a mover behind the scenes and a talent-spotter, mentoring students but not putting his name on their publications, finding grants for those in need, influencing congressional deliberations, shaping science policy, and initiating and co-authoring the textbook *Molecular Biology of the Cell* (1983). Most remarkable of all was his success in appealing to the generous spirit of the 'blue bloods' of Long Island's Gold Coast to support Cold Spring Harbor. His gaffes were overlooked and his snide remarks forgiven, for it was accepted that Watson had remarkable intuition, vision and drive.

The purpose of *Inspiring Science*, the editors write, "is not to reconstruct an academic history or an authoritative biography, but to record friendship and appreciation". Watson requested that it contain "nothing too worshipful and certainly nothing boring". So we are reminded of a number of his less attractive characteristics, notably his intolerance, bad manners and bouts of bad temper.



On the way up: Jim Watson was at Harvard before heading Cold Spring Harbor's biology laboratory.

Cross his path and there will be an explosion; try and steal his staff and you will surely be in for a devastating harangue and lasting resentment. "Jim," writes Norton Zinder, "is the only manager I know who succeeded by pure petulance." In short, he justly earned E. O. Wilson's name for him: the Caligula of biology.

Clearly this book is not a standard scientist's Festschrift. The seven scientific papers in the volume are reprints of Watson's, four of them co-authored by Crick. The structure is chronological, running from his student days to the present, and is supported by an eleven-page historical timeline. For each contributor there is a concise biographical entry to guide the reader. The contributors are former students, collaborators, colleagues (including Wilson, his Harvard opponent), associates in science policy, four members of the Cold Spring Harbor Laboratory's board of trustees, the architect William Grover, and Bruce Alberts and Keith Roberts, co-authors of *Molecular Biology of the Cell*.

Nearly one-third of the book is about the Cold Spring Harbor Laboratory. This section begins with John Cairns' account of his spell as director, when its Biological Laboratory had been amalgamated with the Carnegie Institution's Department of Genetics. He describes how he kept the Biological Laboratory afloat in the 1960s while the scientific board of trustees ran rings around him, making him the fall guy, by holding back the \$25,000 that each of the eight Associated Universities had agreed to contribute to set the laboratory on its feet.

The turning point, according to Cairns, came when Bentley Glass brought the State University of New York (SUNY) into the Associated Universities. Offering to pay at once, SUNY shamed the other institutions into coughing up.

The account by Zinder, a board member from 1967 to 1986, takes us back to the earlier history of the laboratory to understand how it became so run down. It then moves on to Watson's period as director and the subsequent reorganization of the laboratory and its board of trustees.

Building on the successful format of the laboratory's earlier Festschrift to Max Delbrück, *Phage and the Origins of Molecular Biology* (1966), the editors wisely avoided including the much-publicized prelude to the discovery of the structure of DNA. Surprisingly there is no recollection from Ernst Mayr about Watson's Harvard years, nor from anyone in the university administration at that time. And was there no suitable fragment about Cold Spring Harbor from the works of the late Bentley Glass?

Multi-authorship results in differing evaluations of some historical episodes: the impact of Watson's Cold Spring Harbor lecture on DNA in 1953; the trustees' treatment of John Cairns; and the mystery of Watson's effective role as an undergraduate teacher, although, as David Botstein remarks, "the best that can be said of his lectures was that they were unpredictable."

Such differences serve to warn the reader not to take recollections on trust, but to evaluate them critically. If the book had been

presented as an academic history then the lack of exploration of more sources — including archive material from the meetings and correspondence from Harvard's biology department and Cold Spring Harbor's board of trustees — would be a concern. But as a kind of Festschrift, this volume should encourage as well as inform such a work. It will also entertain.

Inspiring Science could be read alongside Victor McElheny's biography of our subject, *Watson and DNA* (for a review see *Nature* 421, 315–316; 2003). The editors are to be congratulated on compiling and organizing such a readable and informative set of recollections that will serve very effectively the role they intended. ■

Robert Olby is in the Department of the History and Philosophy of Science, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA.

An inside view of mental health

The Confinement of the Insane: International Perspectives, 1800–1965

edited by Roy Porter & David Wright
Cambridge University Press: 2003. 371 pp.
£50, \$70

Gerald N. Grob

The history of mental illness and psychiatry before 1960 was written within a Whiggish tradition of inevitable progress, with new psychiatric knowledge and therapies superseding ignorance and barbarism. The subsequent failure of mental hospitals to live up to this promise was attributed to short-sighted and parsimonious public policies that had led to overcrowding and neglect of therapeutic functions.

After 1960, by contrast, the history of psychiatry and mental hospitals underwent fundamental changes. A group of social and psychiatric critics launched an attack on the legitimacy of both psychiatry and mental hospitals and led the development of a quite different synthesis. In so doing they used what purported to be empirical data to support their interpretations. In Michel Foucault's eyes, for example, mental hospitals were institutions of social control that served the interests of a capitalist society that prized productivity above all else. Others, including Thomas Szasz, Erving Goffman and Thomas Scheff, emphasized the repressive nature of such institutions, and some critics even challenged the very existence of mental disorders.

The work of these critics influenced a group of historians who then developed a synthesis that was equally critical of psychiatry and mental hospitals. They emphasized

élite hegemony: people with mental illnesses were simply objects without agency. The confinement of the mentally ill in institutions reflected fears of social disorder and a desire to penalize those who did not contribute to productivity. This scholarship in some ways symbolized the triumph of ideology over reality, as their authors ignored a large body of contradictory data.

In recent years a new and different synthesis has begun to emerge. *The Confinement of the Insane*, a collection of essays by younger scholars that cover developments in no less than fourteen countries on five continents, is but one indication of new trends. Dealing with the institutions and policies of countries with differing populations, traditions and cultures, these scholars largely eschew the angry and polemical writings of the 1960s and 1970s. Basing their analyses on archival data, they present nuanced and subtle interpretations that offer fresh insight into the mental-health policies of different nations.

The authors of these essays avoid a historical analysis that assumes inevitable progress. Yet they find that many of the allegations made by the psychiatric critics lack substance. In their eyes, psychiatry was not a profession characterized solely by self-interest, nor were mental hospitals merely places of confinement for inconvenient people. In place of simplicity, these younger historians offer a complex model that grows out of deep research. They reject the claim that mental hospitals were created by hegemony-seeking élites, and insist that they were multifaceted institutions that simultaneously served caring, therapeutic and custodial functions. Indeed, these institutions were

contested arenas where families, patients, public officials and psychiatrists negotiated among themselves. Admission to the mental hospital, for example, was rarely controlled by police or officials concerned with the maintenance of public order; in most countries, families played key roles in institutionalizing their members.

Moreover, the authors' research undermines older generalizations about the composition of hospital populations. Hospital stays in nineteenth-century England, for example, were relatively brief. In most places, unmarried individuals and males had the highest rates of being institutionalized; females had lower rates, but once inside they tended to remain there longer. The oft-repeated claim that women had disproportionately high rates of incarceration is not supported by empirical data.

Nor were hospitals populated by the fringe elements of industrial society. The analysis of Canadian hospitals, for example, found that most patients were productively employed. Finally, the comparative dimensions of these essays suggests that indigenous cultures and traditions played significant roles in shaping different mental-health policies. These and other data do little to support the claims of ideologically driven scholars that mental hospitals were places where inconvenient and marginal people were incarcerated.

The essays in this volume do not lead to overarching interpretations, but nevertheless they highlight the importance of using patient records and avoiding ideological clichés. Such research can also have contemporary policy implications. Knowledge



Vincent van Gogh painted *The Garden of St Paul's Hospital at St Remy* during a spell in the asylum.

INTERFOTO/BRIDGEMAN ART LIBRARY

Making a case for life

Exhibits at the Natural History Museum in London, such as these models of rat fleas on show in 1927, have been drawing the crowds since the museum opened its doors in 1881. This exhibit was part of a campaign to inform the public of the role of insects in spreading disease.

Photographers, both professional and amateur, have recorded events in the life of 'the world's greatest cathedral to natural history', which has amassed a collection of more than 6,000 images. As well as showing the public face of the museum, the collection contains many pictures showing the minutiae of daily life behind the scenes, as the staff worked to conserve and prepare the exhibits. Arrivals and departures of famous specimens such as the blue whale are recorded, with the style of exhibits changing along with the hairstyles and dress of visitors and staff.

About 100 of these images are themselves now on show, in an exhibition that runs until 20 February 2004 and in a book, *Life Through a Lens* by Susan Snell and Polly Tucker (Natural History Museum Publishing, £15).



NATURAL HISTORY MUSEUM

about the past does not necessarily point the way to future action, but it can serve as an important corrective to misperceptions about the past, which have frequently led to policy-making based on ideology rather than reality. ■

Gerald N. Grob is emeritus professor at the Institute for Health, Health Care Policy, and Aging Research, Rutgers University, 30 College Avenue, New Brunswick, New Jersey 08901, USA.

Core research

The Dynamic Structure of the Deep Earth: An Interdisciplinary Approach

by Shun-ichiro Karato

Princeton University Press: 2003. 264 pp.

\$35, £22.95

Quentin Williams

The deep interior of the Earth is, like space, a seemingly endless source of inspiration for popular fiction. A recent movie, *The Core*, followed this tried-and-tested theme with a tale of travel to the centre of the planet. Along the way, a predictable sequence of characters must die (with their order of demise apparently determined by development, ethnicity, nationality and personality) to ensure that the outrageously attractive protagonists survive to return to the surface.

The realities associated with the deep Earth are actually far more interesting than one might guess from Hollywood's formulaic prosaism. The deep Earth manufactures

diamonds and delivers them, uncombusted, to Earth's surface. Its degassing probably furnished us with Earth's surface water and the starting materials for our atmosphere. And it provides the engine for plate tectonics — the source of mountains, ocean basins and almost everything in between.

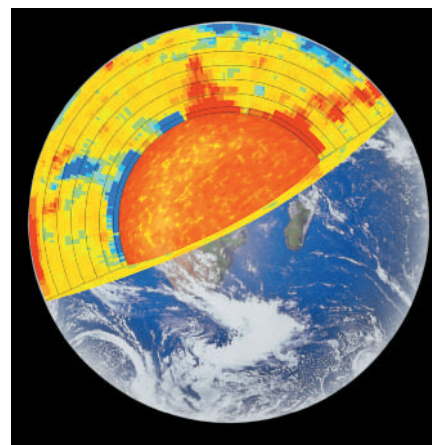
Our understanding of Earth's interior has mushroomed over the past two decades, driven by a synergy between improved seismic imaging, better understanding of the properties of Earth's materials at extreme pressures and temperatures, and computationally intensive simulations of the planet's interior.

The author of *The Dynamic Structure of the Deep Earth*, Shun-ichiro Karato, was trained in rock mechanics and has worked on a broad range of deep-Earth related topics, so he is well equipped to write such a book. Its intent, as outlined in the preface, is not to provide a synopsis, but rather to introduce the interdisciplinary approach of the subtitle through specific problems on which the author has worked. The book seems to belong to a fine tradition of volumes about the deep Earth that incorporate general introductions to research areas but also serve to revisit and remind people of the (mostly already published) viewpoints of the author.

This style perhaps reached its apex with A. E. Ringwood's *Origin of the Earth and Moon* (Springer, 1979), which both placed Ringwood's views into a tutorial framework and provided a focal point for scientific disagreements. This is a lofty benchmark, and the ultimate success of Karato's book will

be determined by the long-term importance and longevity of the views described.

The topics of the book span the depth range of the planet, from the boundary between the rigid outer lithosphere and the underlying fluid asthenosphere to the inner core. The introductory chapter is one of the highlights of the book — a nice pedagogic overview of Earth's structure, complete with boxes containing derivations of the relevant parameters. The author then provides reviews and gives his own take on a selection of current deep-Earth research problems. These include discussions of the lithosphere–asthenosphere boundary and the origin of the 'low-velocity zone' within the shallow upper mantle; the meaning of seismically fast and slow zones in tomographic models of Earth's interior; the relationship between



The rise and fall of slabs of hot (red) and cold (blue) rock in the mantle drives plate tectonics.

S. GRAND, TEXAS UNIV./SPL

Luminet's illuminations

Cosmological modelling and the art of intuition.

Martin Kemp

Letters to *Nature* do not usually open with a peroration on musical harmonics and the organization of the Universe. But Jean-Pierre Luminet, the lead author of a recent paper on the topology of the Universe (*Nature* 425, 593–595; 2003), is no ordinary scientist — at least by today's standards. In some ways, he seems to belong more to the era of Johannes Kepler and his Renaissance predecessors.

Already an advocate of finite models as solutions to cosmological conundrums, he now joins with his colleagues in using measurements of microwave-background temperature fluctuations from the Wilkinson Microwave Anisotropy Probe to propose a model of a finite Universe based on Poincaré dodecahedral space, as already envisaged in Luminet's book *L'Univers Chiffonné* (Fayard, 2001). To convey their vision, they illustrate a beguiling sphere of curved pentagons and a dense view through the surface of a hypersphere tiled with 120 spherical dodecahedra.

To me, a visual historian, the images feel like coming home. After the puzzling 'brane worlds' of Stephen Hawking and others (see *Nature* 415, 738; 2002), we are now only a few multidimensional steps from the model of the cosmos in Kepler's *Mysterium Cosmographicum* (1596), which involves a spherical inscription and concentric nesting and the five regular polyhedra (see *Nature* 393, 123; 1998). Luminet's reinstated visualization of a finite Universe, albeit one from which we can exit through one face and simultaneously enter through the opposite one, relies upon a keplerian form of mental sculpture that may be described as plastic rather than algebraic.

It is not a great surprise to find that Luminet is a co-author, with Marc Lachièze-Rey, of *Celestial Treasury: From the Music of the Spheres to the Conquest of Space* (Cambridge University Press,



Jean-Pierre Luminet's *Big Bang* combines science and art in a multidimensional view of space.

2001), in which such Renaissance luminaries as Kepler and the German artist Albrecht Dürer are given their due. More unexpected are Luminet's activities as an artist and as a published poet. Moreover, he has collaborated with the composer Gérard Grisey on a piece of cosmic music, *Le Noir de l'Etoile*.

Luminet's characteristic lithograph, *Big Bang*, shown here, exploits the spatial vocabulary of perspective to evoke realms beyond the three dimensional. Whereas Escher relied on contradictions and oscillating ambiguity in his graphic art, Luminet suggests plunging, interpenetrating and

vertiginous illogicalities of dynamic space. Spewing from the Big Bang in the top left-hand corner, matter organizes itself into structures on the right; the tumbling dice on the left imply irreversible disorganization arising from chance.

The remarkable range of Luminet's creativity in art and science is integral to his agenda to re-create what he calls a "humanism of knowledge" — not that the arts and sciences are somehow to be conflated, because they work in very different ways, with illogical and logical means. But Luminet argues that they well up from the same instincts and intuitions: "I do not believe that we acquire at the beginning the 'heart of an artist' or the 'heart of a scientist'. There is simply within oneself a single devouring curiosity about the world. This curiosity pushes us to explore it through various languages and modes of expression," he says.

Like science, the arts have developed some fundamental rules: poets have established a series of set forms, for example; painters have pursued variations on perspective; and musicians have developed a harmonic language that is at once simple and intricate. The modes of formal organization have varied over time, yet a work of art habitually relies on tension between set structure and unpredictable freedom, regardless of the kind of formal organization used.

Luminet has the courage to acknowledge the passionate engagement that potentially flows undetected beneath the dry crust of the standard scientific paper. He asks: "Why do we want to separate knowledge from emotion?" His answer is that "amazement about the world is the common root", which is expressed through a "harmonious integration" of all those intellectual and creative faculties that we use to respond to the wonder of things, both immense and minute.

Martin Kemp is professor of the history of art at the University of Oxford and co-director of Wallace Kemp/Artakt.

seismic anisotropy of the mantle and mantle-flow patterns; the rheology of subducted slabs and their geochemical consequences; and possible models for the physical origin of deep earthquakes.

The book concludes with a chapter on core formation, magnetic-field generation and the inner core. This ball of iron at the centre of the Earth has been shown over the past 20 years to have some unusual properties: seismic waves travel faster parallel to the rotational axis than along equatorial paths, and it might (or might not) rotate at a slightly faster rate than the rest of the Earth. The author's own ideas on the origin of anisotropy in the inner core through magnetic-field-induced stresses are described in a somewhat muted way here, leading to a

more balanced feel. The book thus provides a sampler of deep-Earth problems. And as with most samplers, different readers will find nuggets that they like, and others of which they are not so fond.

The key point that emerges, irrespective of the precise flavour of each chapter, is that profound issues of planetary importance about the deep Earth remain unsolved. Conveying this fact achieves a portion of the author's intent in writing this book.

This book was originally published in 2000 in Japanese. This English translation has been substantially updated, but unfortunately the translation is sometimes distracting — some awkward sentences interfere with the flow of the text. Also, a few colour figures would have been worth the invest-

ment — I'd forgotten how horrid seismic tomography looks in shades of grey.

Even with these cosmetic flaws, the book will be of value to practitioners of the deep-Earth sciences. Particularly useful are its handy tutorial derivations of a broad range of simple mathematical concepts, and its reviews of some provocative current topics. I would recommend the book to advanced undergraduates, new graduate students or deep-Earth neophytes, if they are equipped with plenty of complementary reading and are already endowed with what we hope our students develop — a sceptical eye for the deep-Earth guy. ■

Quentin Williams is in the Department of Earth Sciences, University of California, Santa Cruz, California 95064, USA.

Making connections

Summarize yourself in the form of a title of a paper in Nature.

Varied adaptive strategy produces developmental stasis in a paedomorphic acephalate.

What was your first experiment as a child?

I didn't much care for experiments, except mixing chemicals in my chemistry set without following the recipes. I much preferred reading about prehistory and visiting the American Museum of Natural History, where I could lose myself in the exhibits.

Who has been the most important mentor in your career?

Certainly in my undergraduate years, Bob Linsley and Jon Swinchatt at Colgate University in Hamilton, New York. Bob didn't just teach his students, he infected them with his love of palaeontology and evolution. 'Swinch' gave his students the confidence and determination that they needed to pursue their goals — whatever they were. I don't go a day without using something I learned from them. And I've been fortunate to have good mentors throughout my whole career.

What makes a good scientific mentor?

Balance. But achieving it, that's the problem.

What gives you the most job satisfaction now?

What are your major frustrations?

The greatest satisfaction comes from seeing how talented and motivated each new generation of students is. They seem to have boundless energy and ability, and they make it all worthwhile. The greatest frustration is the incredible administrative load of paperwork, regulations and restrictions that increases with each year. It just wastes so much potentially productive time. You'd think that administrators would be making our jobs easier, not more difficult.

What single scientific paper or talk changed your career path?

When I was a graduate student, I read Alan Charig's 1967 paper that tried to explain the evolution of the archosaur pelvis and hindlimb in functional terms. Like most useful papers, its components were eventually superseded by new methods and data (a nice way of saying that most of it was utterly wrong). But it stimulated a lot of young workers to ask new questions in different ways, and it taught me not to accept everything you read but to try to build on it.

What was the worst/most memorable comment you ever received from a referee?

My thesis adviser, Keith Thomson, was brilliant and all his students thought the world

of him. He provided tremendously useful marginal comments on my dissertation chapters. I couldn't read most of them — like my graduate students can't read mine — but that wasn't important. All he needed to do was make a squiggle alongside a passage, and you knew it needed to be fixed. And his detailed comments were kind, frank and perceptive — a tough combination to pull off.

You have the audience in your hands, but some smart-alec asks you the killer question. What's your favourite response?

If you want to keep the audience in your hands, ask the questioner to elaborate. You might learn something. Or he or she could turn out to be a lunatic, which you might put to your advantage.

What book currently resides on your bedside table?

If there's only one book, I'm turning in too early. Currently I've got Robert Richards' *The Romantic Conception of Life: Science and Philosophy in the Age of Goethe*, A.N. Wilson's *The Victorians*, and a couple of pulp westerns and mysteries.

What one thing would you rescue from your burning laboratory?

Specimens that I've borrowed from other museums, because they are literally irreplaceable.

What's the best advice you've ever received?

"Throw the first two strikes high and inside" (Brian Simison); "Check out this new group called The Clash" (Dave Jablonski); "Why don't you apply for that Berkeley job?" (Dave Archibald).

What do you most dislike about having research published?

There's always something to fix or improve when you reread it six months later. But usually my colleagues are kind enough to do it for me.

What would you have become if not a scientist?

It's funny how many prominent colleagues will say privately that they don't think of themselves as competent scientists (and hope no one ever catches on), but at the same time they can't conceive of being anything else. I suppose it would be a fairly cushy job to be an editor for *Nature*...

Is there a tyranny of reductionism in science?

When biology and geology students have to learn physics and chemistry, but the reverse isn't true, we're not valuing the big picture in



Kevin Padian

Kevin Padian is professor of integrative biology and curator in the Museum of Paleontology at the University of California, Berkeley, and president of the National Center for Science Education. His research focuses on how major adaptive evolutionary changes get started.

science education. Mike Bishop, chancellor of the University of California, San Francisco, and a Nobel laureate, stresses the importance of an integrative science education in his recent autobiography. Average people love to learn about stars, Earth and its life. Showing how all the sciences interrelate will help them understand and appreciate molecules and quarks, too.

Name one extravagance you can now get away with because of your eminence.

It's not eminence but advancing age that gives me an excuse for misremembering the names of bright young people that I meet.

What's just around the corner?

Even more amazing revelations in evolutionary developmental biology that will connect genetics, palaeontology and ontogeny. This is the part that Darwin couldn't solve — nor could anyone, until developmental technology caught up with evolutionary questions. We're getting there piece by piece. And the molecules, the fossils and the morphogenetics are making more and more sense together. It's an incredibly exciting time that I've waited for since I was a graduate student.

Whom from the world outside science would you most like to have dinner with?

John Lennon, Pete Townshend and Bruce Springsteen. With a Fender, a Gibson and a Martin nearby.

Do you hope to die before you get old?

Too late!

Expiry dates

Stuart Pimm

The dodo is most certainly dead. But when did the species finally disappear? A statistical approach allows estimation of the date, and could be applied to other extinctions, both past and present.

In 1923, the USS *Tanager* landed a scientific expedition on Laysan, in the northwest Hawaiian Islands. Rabbits, introduced in the late 1800s, had reduced the native vegetation to a wasteland and the island's endemic honeycreeper, the Laysan 'apapane, to a few individuals. During the 1923 visit, a strong storm blew through Laysan, whipping up the dust. When it settled, the honeycreepers were gone, with the time of the species' demise known to the day along with details of the causes of its decline and death. Partly recorded by an early film, the story of the birds' end is unique. Typically, however, we have only a vague idea of when a species goes extinct and the reasons why it did so. On page 245 of this issue, Roberts and Solow¹ provide a method for estimating dates of extinction. Their application is for the most famous recent extinction of all — that of the dodo, discovered on Mauritius then driven to extinction in the seventeenth century.

Roberts and Solow have used a statistical method, involving a formulation known as the Weibull distribution, and have applied it to the record of dodo sightings by sailors and others. Their key result is an estimate of the

probable range of dates for extinction itself, rather than a last sighting, without knowing the underlying statistical distribution of the times observers recorded the species. This result has applications for both past and future extinctions.

Take, for instance, the controversy over whether dinosaurs became extinct on the last and worst day of the Cretaceous, because of the effects of an asteroid impact, or whether the group was in terminal decline millions of years before that. Similar arguments apply to the demise, some 60,000 and 10,000 years ago, of the huge animals that inhabited Australia and the Americas, and of Pacific island bird faunas over the past two millennia. Was the cause a brutal first contact with humans or a consequence of changing climate? For an incomplete fossil record, the last records of species will inevitably precede the date of their extinction, so there will always be a decline in the number of species known to be alive before some posited cataclysm. Extraordinary data have always supplied the key evidence for cataclysms — in the case of the asteroid and the dinosaurs, the iridium layer, seen in rocks, that is thought to have been produced by the

impact. The indictment of humans for recent extinctions gains credibility from narrow windows of human coexistence with their victims. For example, the bones of many bird species barely coincide with charcoal and other features of human presence on Pacific islands.

Roberts and Solow's analysis¹ is applied to *ordinary* data, however — the dates when now-extinct species were recorded. For each species, one could generate the ranges of the predicted date of extinction. Across all the species involved, do these ranges support the hypothesis of a sudden, contingent death, or slow, independent decline, typical of the known background rate of normal extinctions²? The answer could be a mix, of course. Application of Roberts and Solow's method might suggest which dinosaurs (if any) had predicted dates of extinction well before the rest of their kind. In this case, the records are obviously not sightings, as in the case of the dodo, but fossils dated to various times.

The major extinctions now unfolding suggest another application. For well-known species groups, 10% or more are "threatened". That is, expert opinion expects



Figure 1 Humans set about the fauna of Mauritius. In the background, dodos are being bludgeoned to death; in the foreground, a species of parrot (*Lophopsittacus* sp.) is being killed. The latter is thought to have outlived the dodo. But not long afterwards, probably in the early part of

the eighteenth century, this parrot followed its flightless fellow inhabitant of Mauritius down the grim path to extinction. The illustration is by Julian Pender Hume, based on a woodcut from a Dutch journal by Willem van Westzanen.

JULIAN PENDER HUME/NHMP

their demise within a few decades. Birds are the best documented; some 11% are "threatened" and 2% "critically endangered", usually meaning that there is considerable uncertainty about whether the species is still alive. The cover of a catalogue of threatened birds³ depicts two more species from Hawaii — the nukupu'u, seen only sporadically over the past few decades, and the akialoa, not seen in 50 years. The catalogue's title, *Birds to Watch*, implies that there is some hope for these species. Indeed, species sometimes reappear after long absences; the Cebu flowerpecker is one⁴.

There is a problem with retaining hope for species missing in action, however. If we chose the date of extinction to be, say, 50 years after the last record, then we record no extinctions for the past 50 years. Naïve interpretations of counts of the number of extinctions per decade might conclude that extinction rates have declined in the past 50 years. Roberts and Solow's result provides a way of converting records of critically endangered species into predictions of what species might still be found and those for which the expiry date has passed. Other methods, using different statistical approaches individually or in combination, might be applied to the same end.

And what about the dodo — a universal symbol of stupidity and its fatal consequences? The familiar portrait is particularly unflattering, but we forget that so, too, were other bird drawings of that vintage. According to *The New Shorter Oxford English Dictionary*, the dodo "became extinct", perhaps implying that it deserved its fate. Put more accurately, however, the entry would read "humans wrecked its habitat, introduced species that ate it and perhaps directly bludgeoned the flightless birds into oblivion" (Fig. 1) — oblivion most likely coming, at least in the statistical terms of Roberts and Solow's analysis, in 1690, 28 years after the dodo's last confirmed sighting.

A more important question is on what date did we seal its fate? Or that of the Laysan 'apapane? Or that of the Cebu flowerpecker — for its re-discoverers found only three individuals? Exactly when did human actions put these and other species into irrevocable decline? Roberts and Solow do not tackle this much more difficult matter. We must do so, of course, if we are to prevent the extinctions of the large fraction of species now threatened.

Stuart Pimm is at the Nicholas School of the Environment, Duke University, Durham, North Carolina 27708, USA.

e-mail: StuartPimm@aol.com

1. Roberts, D. L. & Solow, A. R. *Nature* **426**, 245 (2003).
2. Lawton, J. H., May, R. M. & Stork, N. in *Extinction Rates* (eds Lawton, J. H. & May, R. M.) 1–24 (Oxford Univ. Press, 1995).
3. Collar, N. J., Crosby, M. J. & Stattersfield, A. J. *Birds to Watch 2: The World List of Threatened Birds*. BirdLife Conservation Ser. No. 4 (BirdLife International, Cambridge, UK, 1994).
4. Magsalay, P., Brooks, T., Dutson, G. & Timmins, R. *Nature* **373**, 294 (1995).

Astronomy

The mystery companion

Noam Soker

A jet-like flow of material, detected in the vicinity of a dying star, supports a model in which such jets shape the gas cloud around the star into a bipolar nebula. The jet probably comes from an unseen companion star.

As Sun-like stars exhaust the nuclear fuel in their cores, they become 'red giants', swelling to a size several hundred times that of the Sun. The stars begin to lose mass, emitting a dense wind of gas that, in most red giants, would be expected to be spherically symmetric. Using the Hubble Space Telescope, Sahai *et al.*¹ have discovered a fast wind emanating from the vicinity of a red giant that is collimated, rather than spherical (page 261 of this issue). This, coupled with similar observations of another dying star², suggests why the gas clouds that form around many red giants — known as planetary nebulae — are elongated, or bipolar. But this latest detection¹, around the star V Hydrae as it begins its transition to a planetary nebula, also points more clearly to the origin of the cloud-shaping jets.

The evolution of a Sun-like star follows a particular cycle. For most of the nuclear-active life of such a star, hydrogen fuses to form helium in the star's core; later, helium fuses to form carbon and oxygen. Once the helium in the core is exhausted, nuclear burning continues in two thin shells surrounding the core: an inner shell of helium fusion to carbon and oxygen, and an outer shell of hydrogen fusion to helium. Many other elements are also formed during this last nuclear burning phase.

The size of the core with its two burning shells is typically only a few per cent of the size of the Sun today. But the star swells. As its production of energy reaches a rate 3,000

times higher than that of the Sun, its size might be up to three times the distance of the Earth from the Sun (hence the Earth may be swallowed by the Sun when our star reaches this final stage of its life, in around 7 billion years). To conserve angular momentum, the rotation of these stars must slow down substantially as their radius increases. So, as the red giants are at this stage losing mass at a very high rate, their slow rotation means that the winds they emit should be spherical. Later in the cycle of evolution, all that remains of the star is a hot bare core, which ionizes and heats its surrounding cloud of gas. This is a planetary nebula.

But most planetary nebulae have bipolar structures, rather than the spherical shape that would be the natural result of their formation from an even stream of gas emission by a red giant. They may be elliptical, or have two opposite narrow lobes or bubbles. This nonspherical structure is a puzzle that astronomers have been struggling to solve for more than 20 years. If the winds from most red giants are spherical, what then is the mechanism that shapes the nonspherical planetary nebulae³?

It has been suggested that jets of matter, thrown out in opposite directions by the dying star, are behind it. Not all astronomers agree — after all, many planetary nebulae are not shaped by jets, so these collimated flows cannot be a universal feature in planetary nebulae. There is also dispute over whether the jets, if they exist, are blown by the giant star

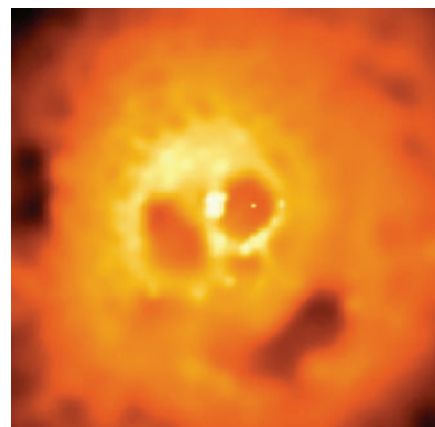


Figure 1 **Blowing bubbles.** At the centre of both the Owl nebula (left) and the Perseus cluster of galaxies (right) is a pair of bubbles, filled with high-temperature, low-density gas — although the bubbles in the galaxy cluster are 100,000 times larger than those in the planetary nebula. In clusters of galaxies, such bubble pairs are known to be formed by two jets. Data reported by Sahai *et al.*¹ suggest that they are also shaped by jets in planetary nebulae. The Owl nebula is shown in optical light; the Perseus image was taken by the Chandra X-ray Observatory.

Plant development

Leaves by number

Fibonacci numbers are notorious for appearing in the most unlikely places, including the architecture of plants. Elsewhere in this issue, Didier Reinhardt and colleagues describe the role of the plant hormone auxin in phyllotaxis, the positioning of leaves around a stem (*Nature* **426**, 255–260; 2003). In so doing, they reveal a mechanism by which Fibonacci numbers can emerge.

Fibonacci was the nickname of the Italian mathematician Leonardo Pisano (1170–1250), who introduced the Arabic number system into Europe. In 1202, he also posed a seemingly inconsequential problem concerning the breeding of rabbits. Its solution required the use of a series of numbers whose members are the sum of the two preceding entries (1, 1, 2, 3, 5, 8, 13, 21, 34, 55, ...). This series now bears his name.

The Fibonacci series occurs in the arrangement of many plant organs. The seeds of sunflowers, *Helianthus annuus*, and the leaves of cacti and succulents (such as *Mammillaria myrtila*, upper image here, and *Sempervivum hybrida*, lower image) are arranged in both left- and right-handed spirals. The numbers of leaves, or seeds, in these spirals are consecutive Fibonacci numbers.

Leaves are also spirally

distributed around the stems of less exotic plants. Here they tend to be separated by an angle of 137.5° . This is the radial equivalent of the golden ratio, ≈ 1.618 , the ultimate proportional increase between successive Fibonacci numbers. Many hypotheses, ranging from the prosaic to the mystical, have been proposed to explain why leaves should stick out at this angle. Some invoke physical properties of the stem, whereas others propose that growing leaves emit an inhibitory field to prevent new leaves from arising in their vicinity, but none has had direct supporting evidence.

Leaves originate at the tips of growing shoots in a self-renewing region known as the shoot apical meristem. Using the thale cress, *Arabidopsis thaliana*, Reinhardt and colleagues saw that new leaves formed where the concentration of auxin was highest. Indeed, artificially adding auxin to specific points on the surface of the meristem caused leaves to be produced at those points.

Auxin is a growth stimulator that is propelled through plant tissues by specific influx and efflux transporters. Reinhardt *et al.* looked at the distribution of one of the most important of these, the efflux protein PIN1, as well as mutants in which it was missing, and deduced that auxin is moved towards the shoot tip



through the outer layers of the shoot apical meristem. Existing leaf buds act as sinks, preventing auxin from continuing its progress directly above them. The maximum auxin concentration, and so the site of new leaf formation, is thus as far away as possible from already-formed leaves.

Reinhardt *et al.* have therefore

uncovered a mechanism in which the position of leaves is determined neither by a physical property of the stem, nor by an inhibitory field produced by growing leaves. Instead, the gaps between leaves, by allowing the free flow of auxin, mark out the position of each new leaf.

Christopher Surridge

itself or by a stellar companion. Increasingly, observations and calculations are pointing to stellar companions as the source of such jets.

Many so-called symbiotic binary systems, which contain a red-giant star and a smaller stellar companion, have bipolar nebulae around them that are remarkably similar to some planetary nebulae. This suggests that binary companions could be at the root of the shaping process in both symbiotic nebulae and planetary nebulae⁴. Some symbiotic systems are known to have jets^{5,6}, and the newly found jets near red-giant stars^{1,2} strongly support jet shaping of some planetary nebulae as well. Sahai and colleagues' discovery¹ is particularly suggestive on this point: the wind speed of the jet from around V Hydrae is much higher than the value expected for material ejected from giant stars; so it seems more likely that the jet is emanating from a much smaller stellar companion as it accretes matter from the

dying star. Furthermore, V Hydrae is known to rotate more rapidly than expected, suggesting that its rate of spin is being affected by a companion star.

The existence of jets in planetary nebulae may also help to explain the similarity between bipolar structures in some planetary nebulae and those in some clusters of galaxies — even though the latter are a million times larger. Both can appear to have oppositely balanced pairs of bubbles in their structure (Fig. 1). In galaxy clusters, these are known to be formed by two oppositely ejected jets. The new finding¹, then, brings us closer to describing a unified formation mechanism for pairs of bubbles in astrophysical systems, over many orders of magnitude in system size.

Stimulating and significant as this jet discovery is, some links are still missing. The collimated flow detected by Sahai *et al.*¹ is less than three years old — is there also a

long-lived jet in that direction, or does this indicate that mass loss is sporadic rather than continuous? A second jet to balance this one would also be expected. As Sahai *et al.* state, the counter-jet may be obscured by dense gas, but it or its signature should now be sought. And, of course, efforts must be made to find the stellar companion, the suspected source of the jets that will shape the planetary nebula of V Hydrae.

Noam Soker is in the Department of Physics, Technion – Israel Institute of Technology, Haifa 32000, Israel.

e-mail: soker@physics.technion.ac.il

1. Sahai, R., Morris, M., Knapp, G. R., Young, K. & Barnbaum, C. *Nature* **426**, 261–264 (2003).
2. Imai, H., Obara, K., Diamond, P. J., Omodaka, T. & Sasao, T. *Nature* **417**, 829–831 (2002).
3. Claussen, M. *Nature* **417**, 798–799 (2002).
4. Corradi, R. L. M. & Schwarz, H. E. *Astron. Astrophys.* **293**, 871–888 (1995).
5. Kellogg, E., Pedely, J. A. & Lyon, R. G. *Astrophys. J.* **563**, L151–L155 (2001).
6. Brocksopp, C. *et al.* *Mon. Not. R. Astron. Soc.* (in the press).

Signal transduction

An eye on organ development

Jonathan A. Epstein and Benjamin G. Neel

Studies in flies and mice have revealed a surprising way in which cells regulate gene activity, with consequences for our understanding of organ formation during development.

The processes by which cells receive outside signals and transmit them into the nucleus have been so well scrutinized that one might think few surprises remain. But three papers in this issue^{1–3} show otherwise. During development, many of the signalling pathways that regulate organ formation are initiated by growth factors that bind to receptors on the cell surface. This triggers a cascade of events inside the cell that frequently involves the addition or removal of phosphate groups (phosphorylation or dephosphorylation, respectively) from cellular proteins. These proteins include those that regulate gene expression, such as transcription factors and their associated ‘co-activators’ and ‘co-repressors’. Rayapureddi *et al.*¹, Tootle *et al.*² and Li *et al.*³ now report the intriguing finding that one type of transcription factor can itself dephosphorylate other proteins — and that this function is crucial in various examples of organ formation. These results reveal a new way in which cells can fine-tune gene expression.

Among the most remarkable discoveries of the past decade has been the identification of single genes that can induce the formation of entire organs or tissues. In fruitflies (*Drosophila melanogaster*), for example, mis-expression of a single gene called *Eyeless*, one of several ‘retinal-determination’ genes that encode transcription factors⁴, is enough to evoke eyes where eyes shouldn’t be (‘ectopic’

eyes). *Eyeless* is a strong inducer of ectopic eyes and is required for the expression of other retinal-determination genes, including *Eyes absent* (*Eya*), *Sine oculis* and *Dachshund*. *Eya* and *Dachshund* can each induce ectopic eye tissue, although only weakly. Expressing *Eya* together with *Sine oculis* or *Dachshund* markedly enhances ectopic eye development, and these genes appear to function together in a molecular network^{5,6}. Consistent with this notion, inactivating any of these genes results in flies lacking eyes⁴. The components of this network have been highly conserved during evolution, with related genes in mammals regulating the development of multiple organ systems. Mutations of the human counterparts cause a variety of congenital disorders⁷.

What do we know about how the products of these genes interact at the molecular level? First, *Sine oculis* and its mammalian relatives are DNA-binding proteins that, depending on the context, can either activate^{8,9} or repress¹⁰ transcription. Second, *Eya* proteins contain highly conserved structural regions known as *Eya* domains, which mediate interactions with *Sine oculis* and *Dachshund*, as well as poorly conserved regions that can activate transcription⁷ (Fig. 1a). Third, when *Eya* and *Sine oculis* are expressed together, they potently activate transcription — even under conditions in which *Sine oculis* alone acts as a repressor^{8–10}. Fourth, *Dachshund* can activate transcription in yeast⁶, but it is also

related to two transcriptional repressors and can bind co-repressors¹⁰.

Now, Rayapureddi *et al.*¹, Tootle *et al.*² and Li *et al.*³ reveal a surprising new function for *Eya* proteins. All three groups started by noting that there is a similarity in amino-acid sequence between the *Eya* domains and sequence ‘signatures’ in enzymes of the haloacid dehalogenase (HAD) superfamily. This large family includes a class of dephosphorylating enzymes (phosphatases)¹¹, some of which remove phosphate groups specifically from serine amino acids in target proteins. Notably, HAD proteins carry out catalysis differently from other well-known phosphatases^{12,13}.

The HAD-related signature in *Eya* domains hinted that the *Eya* proteins might also be phosphatases. Indeed, all three groups show that the *Eya* domains dephosphorylate various artificial substrates. Moreover, mutations in key amino acids within *Eya* domains — including the likely catalytic amino acid, an aspartate — decrease or eliminate catalytic activity. Tootle *et al.*² go on to look at several mutant *Eya* proteins whose phosphatase activity is predicted to be impaired to varying degrees. They find that these proteins are defective in inducing ectopic eyes in flies, and in restoring normal eyes to flies with inactive *Eya*. Rayapureddi *et al.*¹ confirm that mutating the catalytic aspartate of *Eya* results in a decreased (though not entirely missing) ability to restore eyes to flies that lack functional *Eya*.

Li *et al.*³ provide exciting hints as to why phosphatase activity is important for *Eya* function. They generate mice that lack *Six1*, a mammalian relative of *Sine oculis*, and find that these animals have muscle, kidney and skeletal abnormalities, resulting from defective cell proliferation. They go on to show that the genes encoding c-Myc and glial-derived neurotrophic factor are normally repressed by *Six1*. Moreover, this repression is potentiated by *Dach1* (a mouse relative of *Dachshund*), and reversed by *Eya3*. Surprisingly, inactivating *Dach1* by various means also impairs *Six1*-mediated transcription of its target genes. Furthermore, the ability of *Eya3* to reverse repression is destroyed by mutating the catalytic aspartate. Li and colleagues’ conclusion is as unavoidable as it is remarkable: *Dach1* acts as both a co-repressor and a co-activator of *Six1* — and *Eya* proteins, using their embedded phosphatase activities, catalyse the switch from repression to activation (Fig. 1b).

Although there is ample precedent for phosphorylation affecting transcriptional activity, *Eya* proteins represent the first transcription factors with intrinsic phosphatase activity that are capable of modulating transcriptional complexes. Several key questions remain: notably, which proteins does *Eya* dephosphorylate, and how is its activity regulated? In this regard, the reported catalytic

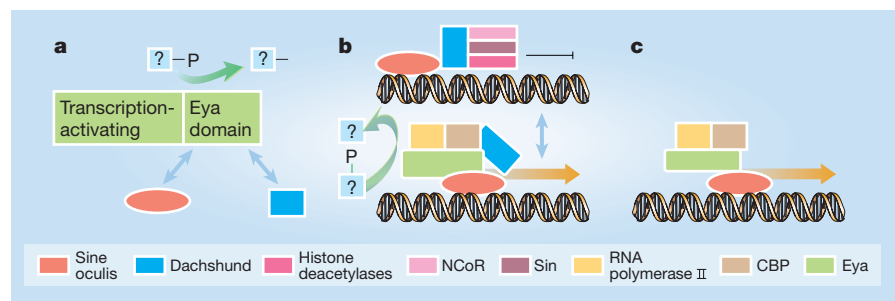


Figure 1 The eyes (absent) have it. **a**, Proteins of the Eyes absent (*Eya*) family are composed of divergent transcription-activating domains and conserved *Eya* domains. The *Eya* domain mediates interactions with the *Sine oculis* and *Dachshund* proteins. The new papers^{1–3} show that the *Eya* domain also works as a phosphatase, removing phosphate groups (‘P’) from tyrosine amino acids (and possibly³ serine and threonine) in unknown target proteins (and possibly² from itself). **b**, **c**, Effects of *Eya* on transcription. **b**, Top, *Dachshund* acts as a transcriptional repressor, which ‘recruits’ co-repressors such as NCoR, Sin and histone deacetylases. Bottom, *Eya*’s phosphatase activity converts *Dachshund* into an activator. The activation complex includes CBP and RNA polymerase II, but the substrate of *Eya*’s phosphatase activity is unknown. **c**, *Eya* and *Sine oculis* together can also activate transcription in a *Dachshund*-independent fashion. This probably does not require phosphatase activity.

properties of the various Eya domains studied by the three groups differ dramatically. For example, the catalytic efficiency of mouse Eya3 varies by more than three orders of magnitude¹⁻³. Moreover, Rayapureddi *et al.* and Tootle *et al.* maintain that Eya3 removes phosphates only from tyrosine amino acids, whereas Li *et al.* claim that it has dual specificity.

What are these target proteins? Tootle *et al.* suggest that Eya3 can remove phosphate groups from itself. Li *et al.* found that Eya3 can dephosphorylate RNA polymerase II — a key transcription enzyme. Further experiments are needed to determine the physiological relevance of both observations. Given that Eya3's phosphatase activity is required to switch Dach1 from a repressor to an activator, another attractive possibility is that Dachshund (or possibly Sine oculis) is an Eya target.

Other differences between the groups' findings may suggest modes of regulation. For example, Tootle *et al.* measured the activity of a synthetic protein containing an Eya domain that is potentially capable of forming a dimer. This protein was much less active than the isolated Eya domain studied by Rayapureddi and colleagues. Dimerization may be of regulatory significance, because Eya can self-associate⁴. As Li *et al.* report, full-length Eya was also considerably less active than the isolated Eya domain, hinting that other domains in the complete protein regulate the Eya domain.

Finally, what is the exact relationship between the catalytic and biological activities of Eya? Tootle *et al.* looked at Eya proteins that had different mutations in the catalytic domain, and found both decreased phosphatase activity and a diminished ability to generate eyes. But there is no clear connection

between the extent of impairment of the two activities. This may reflect details of the experimental design, or, more probably, that Eya has phosphatase-dependent and -independent biological activities (Fig. 1b, c). Alternatively, if an Eya mutant no longer has catalytic activity, but can still associate with its target proteins, then, by sequestering those targets away from potential interacting proteins, it might exhibit some similar effects to the normal phosphatase. Analogous mutants of other phosphatases have such effects¹³. If so, such Eya mutants might provide a rapid means of identifying physiologically relevant targets of these intriguing new phosphatases. ■

Jonathan A. Epstein is in the Cardiovascular Division, Department of Medicine, University of Pennsylvania, 954 BRB II, 421 Curie Boulevard, Philadelphia, Pennsylvania 19104, USA.
e-mail: epsteinj@mail.med.upenn.edu

Benjamin G. Neel is in the Cancer Biology Program, Division of Hematology-Oncology, Department of Medicine, Beth-Israel Deaconess Medical Center, Harvard Medical School, HIM 1047, 330 Brookline Avenue, Boston, Massachusetts 02215, USA.
e-mail: bneel@bidmc.harvard.edu

1. Rayapureddi, J. P. *et al.* *Nature* **426**, 295–298 (2003).
2. Tootle, T. L. *et al.* *Nature* **426**, 299–302 (2003).
3. Li, X. *et al.* *Nature* **426**, 247–254 (2003).
4. Treisman, J. E. *BioEssays* **21**, 843–850 (1999).
5. Pignoni, F. *et al.* *Cell* **91**, 881–891 (1997).
6. Chen, R., Amoui, M., Zhang, Z. & Mardon, G. *Cell* **91**, 883–901 (1997).
7. Wawersik, S. & Maas, R. L. *Hum. Mol. Genet.* **9**, 917–925 (2000).
8. Ohto, H. *et al.* *Mol. Cell. Biol.* **19**, 6815–6824 (1999).
9. Silver, S., Davies, E. L., Doyon, L. & Rebay, I. *Mol. Cell. Biol.* **23**, 5989–5999 (2003).
10. Li, X., Perissi, V., Liu, F., Rose, D. W. & Rosenfeld, M. G. *Science* **297**, 1180–1183 (2002).
11. Aravind, L., Galperin, M. Y. & Koonin, E. G. *Trends Biochem. Sci.* **23**, 127–129 (1998).
12. Barford, D., Das, A. K. & Egloff, M. P. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 133–164 (1998).
13. Tonks, N. K. & Neel, B. G. *Curr. Opin. Cell Biol.* **13**, 182–195 (2001).

Global change

Eruptions linked to El Niño

Shanaka de Silva

Statistical validation of a relationship between explosive volcanic eruptions and the El Niño/Southern Oscillation is a step forward in understanding the effects of such eruptions on climate.

One of the strongest influences on global climate is the cycle of events in the eastern and central tropical Pacific Ocean known as the El Niño/Southern Oscillation (ENSO). The cycle consists of periodic warming and cooling phases, El Niño and La Niña respectively, and conventional wisdom is that they are driven by natural interactions between the atmosphere and ocean in the tropical Pacific¹. On page 274 of this issue, however, Adams *et al.*² present a convincing case that over the past 400 years the ENSO phenomenon has been

influenced by the after-effects of explosive volcanic eruptions. Not only do previous hypotheses about the geophysical processes driving the volcano–ENSO relationship³ now have theoretical and empirical support, but there is new life in an old debate about the role of eruptions in the Earth system.

Although they arise regionally, ENSO events have a worldwide impact because the tropical Pacific is a major heat source that drives atmospheric circulation. Small changes in this heat source, through changes in sea surface temperatures for instance, can



100 YEARS AGO

In the issue of *NATURE* of November 5, Mr. Dixon gave a brief account of some interesting experiments with radium upon seedlings and upon *Volvox*, the results of which were almost entirely negative. Like Mr. Dixon, I have been investigating the action of radium rays upon living matter, but in my experiments animals of simple structure have been employed instead of plants, and my experience leads me to think that the negative result of his experiments may have been due to the distance which separated the small quantity of radium he employed from the seedlings... It is too soon to discuss that obscure problem, the nature of the influence of the rays upon living matter, but it is already clear the experiments with simple forms of life will furnish some data... *Hydra*, both *viridis* and *fusca*, will, as a rule, detach themselves and move out of a pencil of β rays. If, however, the animal is again moved back into the rays from 50 mgr. at 4 mm. distance the third immersion is usually fatal — the tentacles drop off and the body slowly breaks up. Perhaps the most interesting result was obtained with *Euglena viridis*. Encysted specimens under the influence of radium rays (β and γ) in the dark readily become motile and disperse without suffering any harm.

From *Nature* 19 November 1903.

50 YEARS AGO

The work of Prof. F. Zernike, who has been awarded the Nobel Prize for Physics for 1953, includes many distinguished contributions in fields other than optics, but his successes in this branch of physics alone have been sufficient to earn him an international reputation. His serious interest in optics seems to have started some twenty years ago. In 1934 he published a diffraction theory of the Foucault knife-edge test for the figure of astronomical telescope mirrors. This work was notable on two counts: for the discovery of the so-called circle polynomials (now widely, and rightly, known as Zernike polynomials), and for the introduction of the method of phase contrast. Almost immediately Zernike realized that the same method could be used in the other extreme of optical instruments, namely, the microscope. The fruitful results of this invention, especially in biology work, are both sufficiently widely known and appreciated to require no comment.

From *Nature* 21 November 1953.

Figure 1 Blown sky high — a view from the space shuttle *Atlantis* of volcanic dust, high in the Earth's atmosphere, following the 1991 eruption of Mount Pinatubo (below) in the Philippines. Three distinct dust layers are visible.



have widespread consequences — increased snowfall in the Andes, weak hurricane seasons in the Atlantic, drought and associated crop failures in southern Africa, and even wildfires in Indonesia have all been related to ENSO. The climatic effects have repercussions on almost every aspect of human life, in the form of disease outbreaks, low and high agricultural yields, fishery catches, natural disasters and so on.

Stratospheric volcanic aerosols (suspensions of fine liquid or solid particles dispersed in air) are known to influence the absorption, reflection and transmission of solar and terrestrial radiation in the atmosphere. So, given that ENSO is susceptible to small changes in sea surface temperatures, it is natural to consider whether the phenomena are connected. Volcanic aerosols, injected into the stratosphere from explosive tropical

eruptions, result in a complex interplay between stratospheric heating and surface cooling that produces an overall cooling of a few tenths of a degree globally that may last for a few years after the eruption⁴. Regionally, the result may be a change in tropospheric circulation and sea surface temperatures that leads to an ENSO event³.

Support for such a mechanism would come from a demonstration of a statistical connection between the timing of explosive volcanic eruptions and ENSO events. Earlier attempts at identifying such a connection were fraught with problems, most notably in their reliance on data recorded by networks of instruments. These data extend back only into the mid-nineteenth century⁵ and the limited sample size made it impossible to identify a statistically significant correlation between the ENSO cycle and volcanic eruptions. This and other methodological inadequacies have meant that the case for a volcano–ENSO link has remained unsupported⁴.

In their analysis, Adams *et al.*² overcome the deficiencies of the instrumental record by using indices of El Niño and Southern Oscillation activity, derived from palaeoclimate data. They correlate these widely accepted measures of El Niño activity against two well-established independent reconstructions of explosive volcanic activity — the Volcanic Explosivity Index, a measure of the size of the eruption, and the Ice Core Volcanic Index, a measure of potential climate impact. This provides them with a significant dataset of events going back to the mid-seventeenth century. The authors use a standard approach to time-series analysis, the Seasonal Epoch Analysis, and enhance this by ranking, normalizing and preselecting key parameters

to guard against potential bias or undue influence from anomalously large events. Additionally, only the largest eruptions, those most likely to have had an impact on climate, were included in the analysis. This reduces another source of error, in that the timing of the large eruptions is well established and avoids the complication of years such as 1902, when at least five eruptions of varying size occurred. Only one of these is really significant, that of Santa Maria, Guatemala, and it is the only one considered.

The key finding is a recognizable El Niño-like response in years 1, 2 and 3 after large tropical explosive eruptions, followed by a weaker La Niña-like state in years 4, 5 and 6. Adams *et al.* conclude that volcanic eruptions do not trigger all El Niño events but that they help to drive the ocean and atmosphere towards a state in which El Niño-like conditions (warming) are favoured. This is a prudent message, because establishing statistical significance above the noise of the uncertainty in reconstructions of past ENSO behaviour, and the timing and impact of volcanic eruptions, is a tall order: a stronger conclusion is not warranted by their statistical case. Support may come from a better understanding of the evolution of volcanic aerosols by monitoring a large eruption in the future, and from improved models of the interaction between the radiative effects of aerosols and sea surface temperatures in the Pacific.

Nonetheless, Adams and colleagues' work implies that volcanic eruptions, such as that of Mount Pinatubo in 1991 (Fig. 1), may have a larger effect on Earth's climate than previously thought. If they influence the ENSO cycle as proposed, then explosive volcanism is a vital catalyst in global climatic

interconnections, and a major player in Earth's climate system. This is grist to the scientific mill — let debate ensue. ■

Shanaka de Silva is in the Department of Space Studies, University of North Dakota, 526 Clifford Hall, Grand Forks, North Dakota 58202-9008, USA.

e-mail: desilva@space.edu

1. Philander, S. G. H. *El Niño, La Niña and the Southern Oscillation* (Academic, San Diego, 1990).
2. Adams, J. B., Mann, M. E. & Ammann, C. M. *Nature* **426**, 274–278 (2003).
3. Hirono, M. *J. Geophys. Res.* **93**, 5365–5384 (1988).
4. Robock, A. *Rev. Geophys.* **38**, 191–219 (2000).
5. Handler, P. & Andsager, K. *Int. J. Clim.* **10**, 413–424 (1988).

Developmental biology

Gender benders

Peter Koopman

A painstaking triple-gene-knockout study has revealed a crucial role for insulin receptors in male sexual development. But is multiple-gene targeting the way forward for analysing genome function in mammals?

The essential differences between the sexes seem constantly to perplex and beguile us. In mammals such as humans it's all down to the Y chromosome, of course: one either inherits it from one's father or one does not. Some 12 years ago, genetic studies identified what it is about this chromosome that, literally, makes a male. These studies culminated in the discovery of a Y-chromosomal gene known as *Sry*, and provided evidence that it is both necessary and sufficient for the development of testes¹ — the first essential step in producing a male. But just how *Sry* works to bring about the cascade of gene activity leading to testis formation is still not understood. With the paper published on page 291 of this issue by Nef and colleagues², we move a step closer to understanding the complexities of *Sry* and its role in triggering male development.

Nef *et al.* have produced mouse embryos lacking the function of all three members of the insulin-receptor family of genes. These genes are known as *Ir* (this encodes the insulin receptor), *Igf1r* (encoding the receptor for insulin-like growth factor-1) and *Irr* (which produces the insulin-receptor-related receptor), and they are known to regulate various aspects of growth and metabolism. Several studies have suggested that these genes have overlapping functions that could be identified only by inactivating all three at once. Nef *et al.* have done just that. Their triple-gene-knockout embryos resulted from crosses between mice that had mutations in each of the genes individually, produced by experimental gene targeting in embryonic stem cells. None of the three single-knockout mouse strains had defects in sexual development. But the effect of the triple knockout was dramatic and surprising: the embryos with one X and one Y chromosome showed complete male-to-female sex reversal. Their gonads resembled normal XX ovaries in appearance and in the expression of ovary-specific genes such as *Wnt4* and *Fga*. But they lacked the expression of testis marker genes such as *Sox9* and *Amh*.

What does this tell us about the role of the insulin receptors in sexual development? First, the sex reversal was observed very early in the development of the gonads, from the very first time at which it is usually possible to see the difference between testes and ovaries in mouse embryos. So the insulin receptors exert their influence early on.

More telling, perhaps, is that the sex-reversed, triple-knockout embryos showed a dramatic reduction in the expression of *Sry* in the gonads, hinting that the three insulin receptors act upstream of this gene to control sex determination. Some *Sry* gene activity was retained, however, so the receptors are not absolutely required for *Sry* expression. It remains to be seen whether the lower expression levels result from reduced *Sry* activity in each gonadal cell, or whether fewer cells come to express *Sry* but produce it at the usual levels. Another question is how signalling from the three insulin receptors is transduced through gonadal cells to activate *Sry* expression, and how this signalling interacts biochemically and genetically with other factors implicated in *Sry* regulation; these include the product of the tumour-suppressor gene *Wt1* (ref. 3), and the gene-transcription factors GATA4 and FOG2 (ref. 4). And, of course, the six-million-dollar question still eludes us: what happens after *Sry* is activated?

Nef *et al.*² also found that gonadal cells in the triple-knockout embryos proliferated more slowly than their normal counterparts. This observation is significant because an increase in cell proliferation characterizes the early development of XY gonads compared with XX gonads⁵, and is thought to be important for boosting the population of cells that expresses testis-promoting factors, in order to outcompete the ovarian development pathway. So the finding that the insulin receptors are needed for rapid cell proliferation in the gonads supports the idea that the receptors are crucial to the early stages of sex determination. It is not yet clear, though, whether the insulin receptors affect

cell proliferation in the gonads directly, as they do in other organs, or indirectly, via effects on *Sry*; the male-specific increase in proliferation is normally a consequence of *Sry* activity⁵.

The study by Nef and colleagues reveals new and unexpected roles for the insulin-receptor family in sexual development, and goes some way to defining what those functions are. This adds to the known roles for insulin and insulin-like growth factor (Igf), signalling through *Ir* and *Igf1r*, in various metabolic and growth processes. In addition, it provides what may be the first known role for *Irr*, a receptor that is conserved between species but does not bind insulin or Igf molecules. An important goal of future research in sexual development will be to identify the molecule (or molecules) that does bind to *Irr*.

The effect of this triple-gene knockout was dramatic, and the experiment informative. Yet the logistics of generating mutant embryos that lack the function of three genes simultaneously are formidable. For instance, to inactivate a gene completely, both copies of it in the genome must be knocked out — but mice lacking both copies of *Ir* alone die shortly after birth. The same happens with mice lacking *Igf1r* alone. That meant that these mice couldn't be used for breeding to produce triple knockouts. So Nef *et al.* first generated mice that lacked one copy of each of these genes and both copies of *Irr*, and then used these animals for interbreeding to generate the triple-knockout embryos. Generating parents of this genotype would have been trouble enough, but the interbreeding would have produced XY triple-knockout embryos at a frequency of just 1 in 128.

It is clear that single-gene knockouts will become an inadequate tool for comprehensive analysis of gene function in mice, given that more and more gene families are being characterized in which different members act redundantly in the same biological process⁶. But it is also evident that generating multiple mutations by interbreeding of single-knockout animals will not be efficient on a large scale. Arguably, the future of functional gene analysis in complex genetic systems rests on the development of technologies such as RNA interference⁷, which should allow many genes to be disrupted simultaneously *in vivo*. ■

Peter Koopman is at the Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia.

e-mail: p.koopman@imb.uq.edu.au

1. Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. *Nature* **351**, 117–121 (1991).
2. Nef, S. *et al.* *Nature* **426**, 291–295 (2003).
3. Hammes, A. *et al.* *Cell* **106**, 651–659 (2001).
4. Tevosian, S. *et al.* *Development* **129**, 4627–4634 (2002).
5. Schmahl, J., Eicher, E. M., Washburn, L. L. & Capel, B. *Development* **127**, 65–73 (2000).
6. Wellik, D. M. & Capecchi, M. R. *Science* **301**, 363–367 (2003).
7. Xia, H., Mao, Q., Paulson, H. L. & Davidson, B. L. *Nature Biotechnol.* **20**, 1006–1010 (2002).

Medicine

Virus offers life support to cancer

Proc. Natl Acad. Sci. USA
doi:10.1073/pnas.2336099100 (2003)

Many human cancers are causally associated with viruses. For instance, Epstein–Barr virus (EBV) is frequently found in Burkitt's lymphoma — an aggressive type of cancer that affects the immune system's B cells. Now Gregory Kennedy and colleagues find that the viral protein EBNA-1 helps these tumour cells to survive.

One common event that causes Burkitt's lymphoma is a chromosomal translocation that activates *c-myc*, a gene that promotes cell division. Still, most such tumours have additional alterations that affect proliferation and inhibit cell death. In areas where the cancer is endemic, such as Africa, virtually all cases are associated with EBV. But although the virus induces proliferation of infected cells, it's not clear how it supports the tumour.

The only EBV protein consistently expressed in Burkitt's lymphoma cells is a DNA replicator called Epstein–Barr nuclear antigen-1 (EBNA-1). Previously, Kennedy and colleagues constructed an EBNA-1 gene that encodes a non-functional protein, which inhibits its viral counterpart if expressed in the same cell. With this tool to hand, Kennedy *et al.* now show that, without functional EBNA-1, the tumour cells trigger their intrinsic suicide programme and die. Hope springs eternal — although much more research is needed, the finding suggests that EBNA-1 inhibitors might be useful in treating EBV-associated malignancies.

Marie-Thérèse Heemels

Evolutionary biology

Spot the difference

Am. Nat. **162**, 377–389 (2003)

One explanation for the evolution of conspicuous warning signals — for example, as displayed by certain caterpillars (see picture) — is that predators learn to avoid distinctive, 'defended' animals more quickly. Thomas N. Sherratt and Christopher D. Beatty have tested an alternative view, which does not invoke any particular psychological biases in predators. Their experiments involved student hunters and virtual prey; their results support the idea that bright warning colours evolved simply to let the poisonous stand out from the palatable.

Sherratt and Beatty asked students to forage for computer-generated animals, scoring points when attacking an undefended victim, and losing points if they hit a noxious one. The prey varied in their conspicuousness, and passed on their traits



Pretty poisonous: spurge hawkmoth caterpillars (*Hyles euphorbiae*).

to the next generation — if they survived. Defended prey rapidly evolved to become more obvious; even in trials where all prey retained a cryptic appearance, the markings of well-defended and vulnerable animals diverged. Becoming easily detectable might be a good way for inedible animals to distinguish themselves from the rest, the researchers suggest, because an obvious yet edible animal is unlikely to survive for long.

John Whitfield

Organic electronics

Carbon makes contact

Appl. Phys. Lett. **83**, 3945–3947 (2003)

A carbon composite material could help to make electronics all-organic, say Erik J. Brandon and co-workers, by providing the contacts for polymer-based devices. These contacts have previously tended to be made from metals such as gold or platinum, undermining several of the potential advantages of organic devices: ease of processing, low cost and low weight.

The contacts used to 'wire up' the terminals of microelectronic devices must not only be good conductors, but they should also have a work function (the energy needed to remove an electron) matched to the conduction band of the semiconductor from which the device is made. And they have to be chemically stable, sometimes at high voltages. Metals such as gold meet these stringent requirements, but depositing them typically requires high-vacuum conditions.

Brandon and colleagues show that a commercial paste composed of carbon particles dispersed in a heat-curable polyester binder can be used to make contacts about 0.5 mm wide. The contacts have good conductivity, adhesion and flexibility when printed onto silicon dioxide through a stencil, and don't decompose at voltages of up to 100 V. Brandon *et al.* think it should be possible to make much smaller contacts, perhaps using ink-jet printing.

Philip Ball

Atmospheric chemistry

Iron injection

Geophys. Res. Lett. doi:10.1029/2003GL018035 (2003)

Sulphur dioxide, a gas emitted by industrial processes and implicated in acid rain, may be a cloud with a silver lining, according to N. Meskhidze and colleagues. They propose that SO₂ converts iron in mineral dust into a form that can be assimilated as a nutrient by phytoplankton, encouraging primary production in the oceans. As this process 'fixes' atmospheric carbon dioxide in biological tissues, it alleviates global warming.

The limited availability of iron restricts primary production in some regions of the oceans: atmospheric dust is considered to be the main source. But iron in the dust from arid lands is mostly in the form of Fe(III), which is poorly soluble in sea water and thus has low bioavailability. It can be made soluble by acid, and Meskhidze and colleagues think that a prime source of such acid is the SO₂ that dust plumes encounter over urban areas. They confirm that mineral dust transported from the Gobi desert to the Yellow Sea shows a fingerprint of pollutant gases from China.

This doesn't necessarily mean, however, that SO₂ is good on balance for the global climate: the molecules also become oxidized to form sulphate aerosol particles, which have complex effects on the Earth's radiation budget and cloud cover.

Philip Ball

Astronomy

Asteroids warm to analysis

Icarus **166**, 116–130 (2003)

Asteroids are far away and usually just points of light in telescopes. So estimating their size has been a headache for astronomers, especially those who calculate the risk of collision with our planet posed by so-called near-Earth asteroids (NEAs).

Asteroid size is inferred from their albedo, a measure of surface reflectivity which depends on their composition. But the compositions and albedos of NEAs are poorly known, and estimates of NEA masses may be wrong by up to a factor of eight.

Using the Keck I telescope on Mauna Kea in Hawaii, Marco Delbó and colleagues measured the infrared thermal emission from 20 NEAs. From this they deduced the amount of light the NEAs absorb. By subtracting the absorbed light from the known incident sunlight, they inferred values for albedo. The findings increase the number of NEAs for which albedos have been calculated by 35%.

From these results, J. Stuart and one of the paper's co-authors, R. Binzel, produce another estimate: that the Earth should be hit by an asteroid of one kilometre or larger every 600,000 years, a frequency that is some 20–30% lower than previous estimates.

Tom Clarke

Heat reward for insect pollinators

Scarab beetles save on energy by making themselves at home inside a warm flower.

In neotropical forests, adults of many large scarab beetle species spend most of their time inside the floral chambers of heat-producing flowers, where they feed and mate throughout the night and rest during the following day, before briefly flying to another flower. Here we measure floral temperatures in *Philodendron solimoesense* (Araceae) in French Guiana and the respiration rates of *Cyclocephala colasi* beetles at floral and ambient temperatures, and show that the beetles' extra energy requirements for activity are 2.0–4.8 times greater outside the flower than inside it. This finding indicates that heat produced by the flower constitutes an important energy reward to pollinators, allowing them to feed and mate at a fraction of the energy cost that would be required outside the flower.

Self-heating flowers occur in several diverse families of primitive flowering plants in which the female parts are pollinated first, after which pollen is shed¹. Nearly all of these species are pollinated by insects, particularly beetles, which voluntarily remain in the flowers for about 24 hours between the female and male phases^{2,3}.

Thermogenesis by these blooms is likely to increase the dispersal of insect-attracting scents⁴, but this does not explain why heating often continues throughout the period of insect residence, rather than merely during the attraction period⁵; why floral temperature is physiologically maintained in some species at the activity temperature of the insect⁶; and



Figure 1 Warm welcome: the thermogenic flower of *Philodendron solimoesense*, which is found in the neotropical forests of French Guiana, offers a heated chamber as a reward to insect pollinators such as the scarab beetle *Cyclocephala colasi*. The white spathe protrudes from an open scathe; the floral chamber is hidden at the base.

why temperature regulation occurs in the floral chamber and not in the thermogenic, scent-producing organs of some species⁷.

The answers may lie in the fact that many insects require a raised body temperature for activity, which they achieve by metabolically

producing their own heat (endothermy)⁸. Partly because insects are small, the required increase in metabolic rate can be enormous: for example, it increases 16-fold in terrestrially active beetles⁹. A warming floral environment may therefore be a significant energy reward, enabling the insect to reduce the energy cost of its activity.

We measured the effect of floral temperature on energy expenditure by *C. colasi* Endrödi, which is the principal pollinator of *P. solimoesense* A. C. Smith^{10,11} (Fig. 1). Inflorescences showed a rapid increase in spadix temperature after dark, which was associated with strong scent production and the arrival of beetles (Fig. 2a). During the night, when the beetles were still active, the mean floral-chamber temperature was 3.4–5.0 °C warmer than the ambient temperature.

Beetles resting in the respirometer had a low metabolic rate, which increased slightly with rising temperature, whereas the metabolic rate of active beetles was greater and increased with decreasing temperature (Fig. 2b). Active beetles occasionally showed an explosive increase (150-fold) in respiration, indicating an endothermic bout. Such bursts occurred at all temperatures, but were less frequent, shorter and less intense at temperatures above 27 °C.

To estimate more accurately the difference in energy expenditure during periods of rest and of activity, we distributed test temperatures evenly and averaged the metabolic rates over uniform 40-min periods which included shorter bouts of endothermy. Using the temperature of the floral chamber and of the ambient air (Fig. 2a), and applying the energetics equations (Fig. 2b) for active and resting beetles, we calculated the difference in metabolic rate between rest and activity at the two temperatures throughout the night.

The results are expressed as an 'energy-saving factor', which is calculated as the ratio of excess energy required for activity at ambient temperature to that required at the floral temperature (Fig. 2a). This ratio is 4.8 in the evening when the beetles arrive, and decreases to 2.0 at dawn.

Even in the warm lowlands of French Guiana, beetles gain a large energy reward when the floral chamber is only marginally warmer than the air (28 °C and 24 °C, respectively; Fig. 2a). The reward must be even greater for scarab beetles that remain active inside *Philodendron* in the Brazilian highlands, where the air temperature drops to 6 °C (ref. 12). Thermogenesis by flowers pollinated by large scarab beetles is widespread in tropical forests³, with an estimated 900 plant species in at least six families (Cyclan-

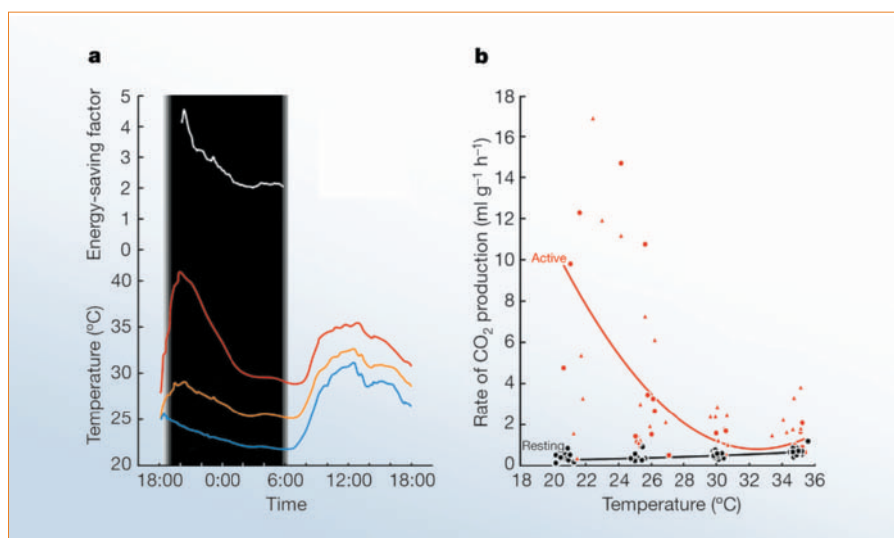


Figure 2 Energetics of *Cyclocephala colasi* beetles in relation to the temperature of *Philodendron solimoesense* flowers. **a**, Temperatures of spadix (red line), floral chamber (orange line) and ambient air (blue line) for 20 inflorescences, measured over 24 h with loggers. The energy-saving factor (white line, top) is calculated as the ratio of the extra energy required for activity at ambient temperature to that required at the temperature of the floral chamber. **b**, Energetics of beetle metabolism, measured as the rate of CO₂ production (V_{CO_2}) by using standard techniques. Results are shown from single beetles (circles) and from groups of three beetles (triangles). The corresponding equations are: resting state (black), $V_{CO_2} = 0.0769e^{0.0616T}$ ($r^2 = 0.42$); active state (red), $V_{CO_2} = 0.06367^T - 4.147 + 68.1$ ($r^2 = 0.40$).

thaceae, Annonaceae, Araceae, Arecaceae, Magnoliaceae and Nymphaeaceae) being visited by more than 220 species of *Cyclocephala* alone¹³. Heat reward may have been even more important during the early evolution of the flowering plants.

Roger S. Seymour*, Craig R. White*,
Marc Gibernau†

*Environmental Biology, University of Adelaide,
Adelaide 5005, Australia

e-mail: roger.seymour@adelaide.edu.au

†Laboratoire d'Évolution et Diversité Biologique,
Université Paul Sabatier, 31062 Toulouse, France

1. Thien, L. B., Azuma, H. & Kawano, S. *Int. J. Plant Sci.* 161 (suppl.), 225–235 (2000).
2. Gottsberger, G. *Taxon* 37, 630–643 (1988).
3. Bernhardt, P. *Plant Syst. Evol.* 222, 293–320 (2000).
4. Endress, P. K. *Diversity and Evolutionary Biology of Tropical Flowers* (Cambridge Univ. Press, Cambridge, 1994).
5. Seymour, R. S. *J. Exp. Bot.* 50, 845–852 (1999).
6. Seymour, R. S. & Schultze-Motel, P. *Endeavour* 21, 125–129 (1997).
7. Seymour, R. S. & Schultze-Motel, P. *Proc. R. Soc. Lond. B* 266, 1975–1983 (1999).
8. Heinrich, B. *The Hot-blooded Insects: Strategies and Mechanisms of Thermoregulation* (Harvard Univ. Press, Cambridge, Massachusetts, 1993).
9. Bartholomew, G. A. & Casey, T. M. *Science* 195, 882–883 (1977).
10. Gibernau, M. & Barabé, D. *Can. J. Bot.* 78, 685–689 (2000).
11. Barabé, D., Gibernau, M. & Forest, F. *Bot. J. Linn. Soc.* 139, 79–86 (2002).
12. Gottsberger, G. & Silberbauer-Gottsberger, I. *Biotropica* 23, 23–28 (1991).
13. Schatz, G. E. in *Reproductive Ecology of Tropical Forest Plants* (eds Bawa, K. S. & Hadley, M.) 69–84 (UNESCO, Paris, 1990).

Competing financial interests: declared none.

Psychophysics

How fielders arrive in time to catch the ball

Tracking an object moving in three dimensions, whether as an insect pursuing a mate on the wing¹ or as a batsman aiming to hit an approaching ball², provides the spatial and temporal information needed to intercept it. Here we show how fielders use such tracking signals to arrive at the right place in time to catch a ball — they run so that their angle of gaze elevation to the ball increases at a decreasing rate while their horizontal gaze angle to the ball increases at a constant rate (unless the distance to be run is small). Allowing the horizontal angle to increase minimizes the acceleration that the fielder must achieve to reach the interception point at the same time as the ball³.

Figure 1a shows two sources of information that the fielder receives from tracking the ball: the angle of elevation of gaze (α) and the horizontal angle through which the gaze system has rotated relative to the initial direction from fielder to ball (δ). Fielders run so that α increases at a decreasing rate⁴ — in principle, this guarantees interception. The locus of points from which α has increased by the appropriate amount in the ensuing time interval lie on a circle centred on the vertical projection of the ball to the ground. As the

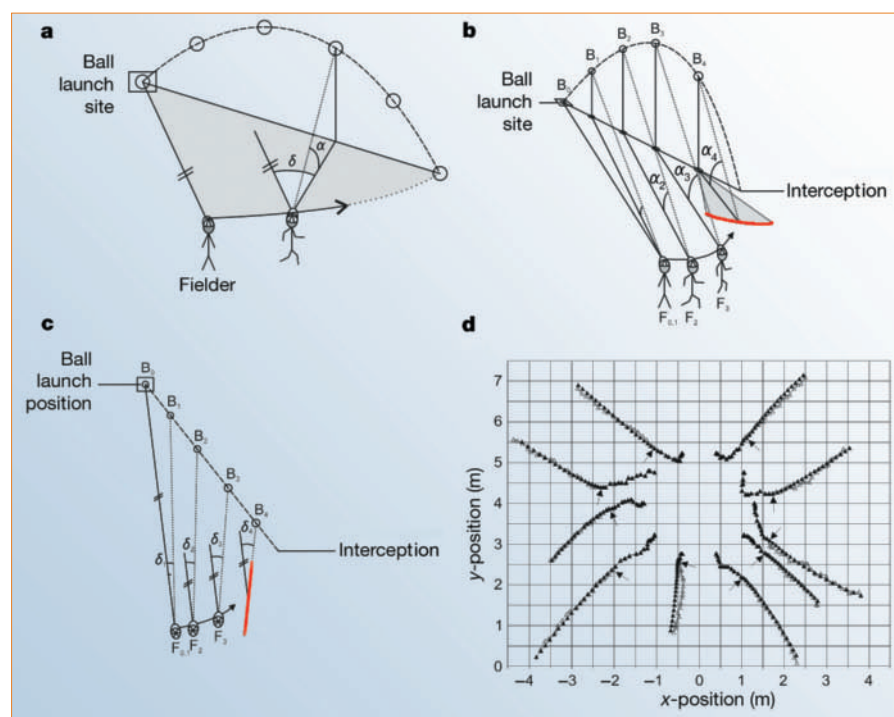


Figure 1 Keeping an eye on the ball. **a**, Angles of gaze from fielder to ball: α is the angle of gaze elevation; δ is the angle of horizontal gaze rotation relative to starting direction. **b**, Movement of a fielder to adjust α . As indicated by the arrow, a fielder at position F_3 must move to a position on the red arc bounding the shaded region in the ensuing time interval to ensure that $d\alpha/dt$ decreases at a particular rate. **c**, Aerial view of ball and fielder; the red line shows the locus of points to which the fielder at position F_3 must move in the ensuing time interval to ensure that $d\delta/dt$ remains constant. **d**, Paths of real (light triangles) and simulated (dark triangles) fielders for ten catches. Initially, the simulated fielder was moved to the same positions as the real fielder; after the point marked by the arrow, it moved under the constraints for α and δ described in the text.

ball falls, the circles decrease in diameter and the fielder, moving through successive circles, homes in on the interception point (Fig. 1b).

However, running so that α increases at a decreasing rate would not in itself be an effective strategy. Early in the ball's flight, there is little constraint on the direction in which the fielder must run, as the circle of points that satisfies the α constraint is large. If the fielder does not initially move towards the interception point, he may not be able to run fast enough to implement the strategy as the flight nears its end. Efficient interception requires the fielder to select a point in the circle that is towards the interception point⁵.

We have found that fielders usually run so that δ increases at a constant rate³. This requires them to move at successive time intervals to a point from which δ would have an appropriate value, given the position of the ball at the end of the time interval. These points lie on a straight line (Fig. 1c). To satisfy both constraints, the fielder must move to the point at which the line and circle intersect, thereby ensuring that the fielder is moving towards the interception point throughout the ball's flight.

We simulated a fielder moving under these constraints to catch balls on the same trajectories as those caught by real fielders. The simulated fielder was moved to the same positions as the real fielder for the first 600 ms to give the same initial visual experience

as the fielder, and then moved so that α increased at the average declining rate and δ increased at the average rate experienced by the real fielder up to that point. The similarity between the real and simulated fielders' running paths can be seen in Fig. 1d.

According to our theory, fielders do not explicitly select their direction or speed when running. They move in way that satisfies the constraints for the changes in α and δ . This view of the fielder's strategy as based on repeated local constraint satisfaction, rather than on a calculation of where the ball will ultimately fall, is consistent with the subjective sensation of running to catch a ball. You do not know exactly where the ball will land, but you do know whether or not you will be able to intercept it. This presumably reflects the knowledge that you have been able to satisfy the constraints (or not).

The origin of the strategy may be that a child watching objects that hit him or her will experience α increasing at a decreasing rate, whereas watching balls that pass by will produce a declining α and an accelerating δ (ref. 3). When the child tries to catch the ball, he or she will run in a way that reproduces previous experiences with objects that hit and avoids those from objects that missed.

Peter McLeod*, Nick Reed*, Zoltan Dienes†

*Department of Experimental Psychology,
University of Oxford, South Parks Road,
Oxford OX1 3UD, UK

e-mail: peter.mcleod@psy.ox.ac.uk

†Department of Experimental Psychology,
University of Sussex, Falmer,
Brighton BN1 9QG, UK

1. Land, M. F. & Collett, T. S. *J. Comp. Physiol.* **89**, 331–357 (1974).
2. Land, M. F. & McLeod, P. *Nature Neurosci.* **3**, 1340–1345 (2000).
3. McLeod, P., Reed, N. & Dienes, Z. *J. Exp. Psychol. Hum. Percept. Perf.* (submitted).
4. McLeod, P., Reed, N. & Dienes, Z. *J. Exp. Psychol. Hum. Percept. Perf.* **27**, 1347–1355 (2001).
5. Tresilian, J. Q. *J. Exp. Psychol.* **48**, 688–715 (1995).

Competing financial interests: declared none.

Flightless birds

When did the dodo become extinct?

The extinction of the dodo (*Raphus cucullatus* L.; Fig. 1) is commonly dated to the last confirmed sighting in 1662, reported by Volkert Evertsz on an islet off Mauritius^{1,2}. By this time, the dodo had become extremely rare — the previous sighting having been 24 years earlier — but the species probably persisted unseen beyond this date. Here we use a statistical method to establish the actual extinction time of the dodo as 1690, almost 30 years after its most recent sighting.

In most cases, the extinction of a species must be inferred from the record of sightings or from collections of individual organisms. But when a species becomes increasingly rare before its final extinction, it may continue to exist unseen for many years — so the time of its last sighting may be a poor estimate of the time of extinction.

We applied an optimal linear estimation method based on the sighting record of the

dodo to determine when the bird finally became extinct. Let $T_1 > T_2 > \dots > T_k$ be the k most recent sighting times of a species, ordered from most recent to least recent. Interest centres on using this record to estimate the extinction time, θ . In this context, optimal linear estimation is based on the remarkable result that the joint distribution of the k most recent sighting times has (at least roughly) the same ‘Weibull form’, regardless of the parent distribution of the complete sighting record³.

Briefly, the optimal linear estimator of θ has the form of a weighted sum of the sighting times, calculated as

$$\hat{\theta} = \sum_{i=1}^k a_i T_i$$

The vector of weights is given by

$$a = (e^T \Lambda^{-1} e)^{-1} \Lambda^{-1} e$$

where e is a vector of k 1’s and Λ is the symmetric $k \times k$ matrix with typical element $\lambda_{ij} = (\Gamma(2\hat{\nu} + i)\Gamma(\hat{\nu} + j))/(\Gamma(\hat{\nu} + i)\Gamma(j))$, $j \leq i$, and where Γ is the standard gamma function. Also,

$$\hat{\nu} = \frac{1}{k-1} \sum_{i=1}^{k-2} \log \frac{T_i - T_k}{T_i - T_{i+1}}$$

is an estimate of the shape parameter of the joint Weibull distribution of the k most recent sighting times. An approximate $1 - \alpha$ confidence interval for θ is given by

$$(T_1 + \frac{T_1 - T_k}{S_L - 1}, T_1 + \frac{T_1 - T_k}{S_U - 1})$$

where $S_L = (-\log(1 - \alpha/2)/k)^{-\hat{\nu}}$ and $S_U = (-\log(\alpha/2)/k)^{-\hat{\nu}}$.

The $k=10$ most recent confirmed sight-

ing times of the dodo are 1662, 1638, 1631, 1628, 1628, 1611, 1607, 1602, 1601 and 1598.

An escaped slave named Simon claimed to have seen a dodo as recently as 1674. However, the reliability of this and other later claims are open to question. For this record, the estimated shape parameter is $\hat{\nu} = 0.39$ and the estimated extinction time is $\hat{\theta} = 1690$. The approximate 0.95 confidence interval for θ is (1669, 1797). The width of this interval is a result of the low sighting rate of the dodo at the end of its sighting record. Because this rate was so low, it is impossible to rule out an extinction date as late as 1797. This implies that the purported sighting in 1674 cannot be ruled out on the basis of extinction time alone. If a sighting in 1674 is included in the record, the estimated extinction time is extended only modestly to 1700 and actually narrows the confidence interval to (1679, 1790).

David L. Roberts*, Andrew R. Solow†

*Royal Botanic Gardens, Kew, Richmond,
Surrey TW9 3AE, UK

†Woods Hole Oceanographic Institution,
Woods Hole, Massachusetts 02543, USA

e-mail: asolow@whoi.edu

1. Cheke, A. S. in *Studies of Mascarene Island Birds* (ed. Diamond, A. W.) 5–89 (Cambridge Univ. Press, Cambridge, 1987).
2. Fuller, E. *Dodo: A Brief History* (Universe Books, New York, 2002).
3. Cooke, P. *Biometrika* **67**, 257–258 (1980).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Astronomy

Black holes, fleas and microlithography

Fresnel lenses allow almost perfect imaging in widely different circumstances, but their focus is perfect only for a single wavelength. Wang *et al.*¹ have shown how the effective bandpass may be widened for X-ray microscopy by using a compound diffractive/refractive lens near to an absorption edge. A compound lens has also been proposed for high-energy astronomy, working well above all absorption edges^{2,3}. Although the scale is very different, we point out here that the principle is the same. Ever since Galileo constructed an astronomical telescope that he was able to reconfigure to study fleas and gnats, astronomy and microscopy have relied on optics that are closely related, but different in detail.

In principle, the highest resolution in each case is obtained with the shortest wavelengths — X-rays or γ -rays. In microscopy and microlithography, the spatial resolution of the optical system dictates the level of detail in images and the fineness of structures in integrated circuits. Resolution better than 30 nm has been reported⁴

Figure 1 Dead as a dodo: the flightless bird from Mauritius and the adjacent islands weighed in at about 23 kg and was hunted to extinction. Its last confirmed sighting was in 1662, although an escaped slave claimed to have seen the bird as recently as 1674. In fact, it is estimated by using a Weibull distribution method that the dodo may have persisted until 1690, almost 30 years after its presumed extinction date. Although gone forever, the dodo’s lumbering appearance in Lewis Carroll’s *Alice’s Adventures in Wonderland* has ensured that it will not be forgotten.



— a value that is close to the theoretical limit.

In astronomy, it is angular resolution that is important. For X-rays, the best available angular resolution (0.5 arc seconds (arcsec) with the Chandra observatory³) is about 10^4 times worse than the diffraction limit. There are good reasons for trying to do better: attaining one of astronomy's 'holy grails' — to 'see' a black hole by imaging the space-time around the supermassive objects at the centres of active galaxies — would require resolution below 1 micro-arcsec, corresponding to the diffraction limit for a 60-cm-diameter system using 500-keV γ -rays, or a 60-m-diameter one using 5-keV X-rays.

In microscopy, the best resolution is obtained using Fresnel zone plates and the closely related phase Fresnel lenses. These are highly chromatic, with a focal length that is inversely proportional to wavelength, so the best resolution is obtained with essentially monochromatic radiation. Wang *et al.*¹ have proposed a scheme for making an achromatic X-ray lens by combining a Fresnel lens with a refractive component (Fig. 1). Such lenses should have the same high resolution but with a larger useful bandwidth.

Fresnel lenses have also been considered for high-energy astronomy^{2,3,6}, and the achromatic Fresnel lens scheme suggested by Wang *et al.* has also been proposed in that context³. Figure 1 shows a general achromatic lens scheme of this type. It has even been shown (L. Speybroeck, unpublished work, and ref. 6) that the second derivative of focal length, as well as the first, can be corrected by separating the two components — a configuration that may also find application in microscopy.

The achromatic configuration proposed for microscopy differs in one detail. Wang *et al.* suggest taking advantage of the high dispersion that occurs close to X-ray absorption edges. Over wider bands and for short wavelengths, λ , the dispersion, D , that they define is close to zero, but correction remains possible because of the λ^{-2} component in

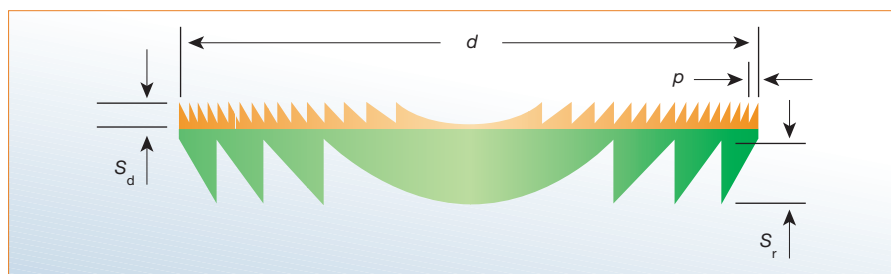


Figure 1 An achromatic refractive/diffractive lens. The diffractive component (orange) could be a phase Fresnel lens (as shown) or a (phase) zone plate. The refractive component (green) is stepped to avoid excessive absorption. Its curvature is exaggerated here for clarity. The same principle can be used both in astronomy³ and in microscopy/microlithography¹, but with different parameters (Table 1). In the microscope objective case, stepping the refractive component introduces no error; in the astronomical case, some resolution is lost, but multiple passbands increase the throughput. d , Diameter of the lens; S_d and S_r are the depth of the steps in the diffractive and refractive components, respectively, and correspond to phase shifts of an integer number of cycles at the nominal wavelength; p is the finest pitch of the diffractive element.

the expression for the focal length, f . The radii of curvature of the refractive components are much less favourable, but in astronomy the considerations are different. Spatial resolution, which is of interest in microscopy, depends on $\lambda f/d$, where d is the diameter of the lens, whereas angular resolution depends only on λ/d . Thus, long focal lengths can be used for astronomy, minimizing both the chromatic aberration that requires correction and the curvature of the components needed to correct it. Furthermore, with large d and short λ , micro-arcsec resolution does not necessarily require perfect phase coherence.

Building a micro-arcsec telescope based on a Fresnel lens will not be easy. The focal length could be up to a million kilometres, and two or three separate satellites carrying the lens(es) and the focal plane array will have to be positioned with centimetre accuracy. But micro-arcsec imaging is an important target (NASA recently included a black-hole imager in its SEUS Roadmap⁷) and other approaches to achieving this goal, such as an X-ray interferometer carried on a flotilla of spacecraft^{8,9}, have requirements that are just as extreme, if not more so.

Although overall system performance, including spacecraft positioning for example, will have to be taken into account, for

Fresnel-lens-based telescopes the purely optical performance can be predicted with confidence. The basic physical principle has already been demonstrated in microscopy, where achromatic lenses will presumably also soon come into use. The different requirements of astronomy merely require a change of scale: lenses of many metres in diameter with millimetre-scale zones, instead of sub-millimetre lenses with zones of the order of 50 nm (Table 1). Constructional tolerances will be correspondingly easier to achieve.

An achromatic Fresnel-lens telescope measuring 10^{11} m in length will certainly not be built very soon. Perhaps experience with lenses that are limited by diffraction at the nanometre scale will give us the confidence to proceed to imaging black holes with millions of times the mass of our Sun.

Gerry Skinner*, Paul Gorenstein†

*Centre d'Étude Spatiale des Rayonnements, Toulouse 31208, France

e-mail: skinner@cesr.fr

†Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA

- Wang, Y., Yun, W. & Jacobsen, C. *Nature* **424**, 50–53 (2003).
- Skinner, G. K. *Astron. Astrophys.* **375**, 691–700 (2001).
- Skinner, G. K. *Astron. Astrophys.* **383**, 352–359 (2002).
- Gil, D., Menon, R. & Smith, H. I. *J. Vac. Sci. Tech. B* (in the press).
- Jerius, D. *et al. Proc. SPIE* **4012**, 17–27 (2000).
- Gorenstein, P. *Proc. SPIE* **4851**, 599–606 (2003).
- NASA Structure and Evolution of the Universe Subcommittee. *Beyond Einstein: From the Big Bang to Black Holes* (NASA, 2003).
- White, N. *Nature* **407**, 146–147 (2000).
- Cash, W. *Proc. SPIE* **4852**, 196–209 (2003).

erratum

Mechanism for 'superluminal' tunnelling

Herbert G. Winful

Nature **424**, 638 (2003)

The following information was omitted from Fig. 1 of this communication: time is in units of the transit time, L/c , through an equivalent distance *in vacuo*; the brown curve is the incident pulse and the blue curve represents the tunneled pulse; also, the tunneled pulse has been scaled by a factor of $1/0.0014$ to make it visible.

Table 1 Lens parameters for astronomy and microscopy/microlithography

	Astronomy ³	Microscopy ¹
Lens diameter	5 m	1 mm
Focal length at nominal wavelength	+40,000 km	+22.475 mm
Wavelength, λ (photon energy)	0.062 nm (20 keV)	1.335 nm (929 eV)
Theoretical resolution	5×10^{-6} arcsec	35 nm
Diffractive component		
Pitch at edge	0.5 mm	60 nm
Focal length	+20,000 km	+22.475 mm
Refractive component		
Material	Polycarbonate	Copper
Focal length	–40,000 km	–
Maximum thickness*	20 mm	<1 μ m

Apart from the obviously disparate scales, the differences are driven by the dispersion $Z = (d\delta/d\lambda)(\lambda/\delta)$, where $\delta = 1 - \mu$ and μ is the refractive index. We adopt this measure of dispersion instead of the parameter $D = df/d\lambda$ used in ref. 1 as it includes the dispersion present even at short wavelengths, where the atomic scattering factor f_i is constant. Well above absorption edges (astronomy), Z is very close to -2 , independent of the material. In the band close to the copper L absorption edge (microscopy/microlithography), Z is large and varies in the range -800 to $+250$.

*Based on 50% mean transmission.

Eya protein phosphatase activity regulates Six1–Dach–Eya transcriptional effects in mammalian organogenesis

Xue Li¹, Kenneth A. Oghi¹, Jie Zhang¹, Anna Krones¹, Kevin T. Bush², Christopher K. Glass³, Sanjay K. Nigam², Aneel K. Aggarwal⁴, Richard Maas⁵, David W. Rose⁶ & Michael G. Rosenfeld¹

¹Howard Hughes Medical Institute, School and Department of Medicine, UCSD, 9500 Gilman Drive, Room 345, La Jolla, California 92093-0648, USA

²Departments of Medicine and Pediatrics, School and Department of Medicine, UCSD, 9500 Gilman Drive, La Jolla, California 92093-0693, USA

³Department of Cellular and Molecular Medicine, School and Department of Medicine, UCSD, 9500 Gilman Drive, La Jolla, California 92093-0651, USA

⁴Structural Biology Program, Department of Physiology and Biophysics, Mount Sinai School of Medicine, Box 1677, 1425 Madison Avenue, New York, New York 10029-6574, USA

⁵Department of Medicine/Division of Genetics, Brigham and Women's Hospital/Harvard Medical School, 20 Shattuck Street, Thorn 1010 Boston, Massachusetts 02115, USA

⁶Department of Medicine, University of California, San Diego 9500, Gilman Drive, La Jolla, California 92093-0673, USA

The precise mechanistic relationship between gene activation and repression events is a central question in mammalian organogenesis, as exemplified by the evolutionarily conserved *sine oculis* (Six), *eyes absent* (Eya) and *dachshund* (Dach) network of genetically interacting proteins. Here, we report that Six1 is required for the development of murine kidney, muscle and inner ear, and that it exhibits synergistic genetic interactions with Eya factors. We demonstrate that the Eya family has a protein phosphatase function, and that its enzymatic activity is required for regulating genes encoding growth control and signalling molecules, modulating precursor cell proliferation. The phosphatase function of Eya switches the function of Six1–Dach from repression to activation, causing transcriptional activation through recruitment of co-activators. The gene-specific recruitment of a co-activator with intrinsic phosphatase activity provides a molecular mechanism for activation of specific gene targets, including those regulating precursor cell proliferation and survival in mammalian organogenesis.

Transcriptional repression and activation of genes control essential cellular functions including cell proliferation, differentiation and cell death during organogenesis. Genetic studies in *Drosophila*, for instance, have identified a synergistic nuclear complex consisting of *sine oculis* (*so*) DNA-binding homeodomain factor¹ and *eyes absent* (*eya*)² and *dachshund* (*dac*)³ nuclear cofactors, mutations of which lead to the failure of eye formation, and ectopic expression of which leads to additional eye formation^{4,5}. The highly conserved vertebrate homologues Six1–6 (ref. 6), Eya1–4 (ref. 7) and Dach1–2/Ski/Sno (ref. 8) respectively, are co-expressed in multiple organs, including eye, inner ear, pituitary gland, muscle and kidney, and thereby provide an excellent model system to study transcriptional regulatory mechanisms during organogenesis. Indeed, Dach2, Eya2 and Six1 synergistically regulate myogenesis in chicken somite culture⁹.

Both Eya and Dach have been proposed to function as cofactors for Six, and genetic studies in *Drosophila* have demonstrated synergistic interactions between *so*, *eya* and *dac* during eye development^{10,11}. Eya has no apparent DNA-binding activity and is translocated from the cytoplasm to the nucleus by Six proteins, and serves as a co-activator of Six in the regulation of downstream genes¹². In contrast, Dach is closely related to Ski/Sno^{8,13}, which are known transcription co-repressors that form strong repression complexes with N-CoR, Sin3A and members of histone deacetylases in order to functionally repress downstream gene expression¹⁴.

Six6, together with Dach, represses the cyclin-dependent protein kinase inhibitor *p27^{kip1}*, thus controlling retinal and pituitary precursor cell proliferation¹⁴. Six3, which is closely related to Six6, also functions as a transcriptional repressor, interacting with the Groucho/Tle co-repressor families¹⁵, and it is essential for eye and

other rostral structure development in mice by means of direct repression of Wnt genes¹⁶. Thus, through interactions with specific co-repressors, Six3 and Six6 function as evolutionarily conserved tissue-specific transcriptional repressors required for precursor cell proliferation and differentiation.

Here we demonstrate genetic interactions between Six and Eya factors that regulate precursor cell proliferation and survival during mammalian organogenesis. Our data provide initial evidence for gene-specific recruitment of a co-activator (Eya) with essential intrinsic protein phosphatase activity that is required for regulation of specific gene targets controlling precursor cell proliferation and survival during mammalian organogenesis.

Six1 regulates precursor cell proliferation and survival

On the basis of the actions of Six6 and Six3, we investigated the potential role of Six1 in the target tissues in which it is highly expressed, including kidney, otic placode, skeletal muscle, pituitary and developing nasal structures. This was achieved by generating mice null for the *Six1* genomic locus using homologous recombination in embryonic stem cells by replacing the functionally conserved Six domain (SD) and DNA-binding homeodomain (HD) with the IRES-LacZ and PGK-Neo sequences¹⁴. This results in loss of *Six1* transcripts and expression of *LacZ*, which recapitulates the endogenous *Six1* expression pattern (Fig. 1a; see also Supplementary Fig. 1). Profound effects were noted in virtually all organs in which Six1 is expressed, including a failure of renal organogenesis with variable penetrance; that is, from virtual absence of kidneys to a marked, often asymmetrical reduction of kidney size (Fig. 1b). Immunohistochemical analyses of the rudimentary kidneys from

Six1^{-/-} mutants revealed no obvious aberrancy of the branching process¹⁷ (data not shown). Muscle development was also profoundly affected, exemplified by a marked reduction in 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) staining of forelimb bud (Fig. 1c). Indeed, most migratory hypaxial muscles, including those of forelimbs, the diaphragm and the tongue were missing in mutant embryos (Figs 1d and 2a). Hindlimb muscles were less affected, especially at the proximal region (see below, and data not shown). All early markers of muscle development (see below) appeared normally, in accordance with an independent analysis¹⁸, but the number of potential *Six1*-expressing cells (LacZ⁺) was markedly decreased (Figs 1c, 2a and 3d). There was also a defect of nasal development, but despite strong expression of *Six1*, the anterior pituitary showed only a minimal (<10%) decrease in size at low penetrance in *Six1*^{-/-} mice (see below, and data not shown). In

addition, there was invariably severe rib-cage deformation, often with the fusing of distal rib cartilage (Fig. 1e, panels 1–4) and virtual loss of the inner ear structures (Fig. 1e, panels 5–8).

On the basis of these data it became important to define the molecular basis of *Six1*-induced alterations in target tissue development. We therefore analysed expression of key markers, including *Pax3* (ref. 19), *Lbx1* (refs 20–22), *c-Met* and *Hgf* (ref. 23), and the results demonstrated normal cellular expression of all of these markers but severe reduction of muscle precursor cells, consistent with the X-gal staining pattern. Notably, the expression level of *Gdnf*²⁴ and *Six2* (ref. 25) per cell was reduced compared with a heterozygous littermate (Fig. 2a). Marker genes, including *Eya1* (ref. 25), *Pax2* (ref. 26), *c-Ret*²⁷ and *Wt1* (ref. 28), as well as *Wnt* genes²⁹, displayed normal cellular expression during early kidney organogenesis at embryonic day (E)10.5 and E11.0 (ref. 30) (Fig. 2b,

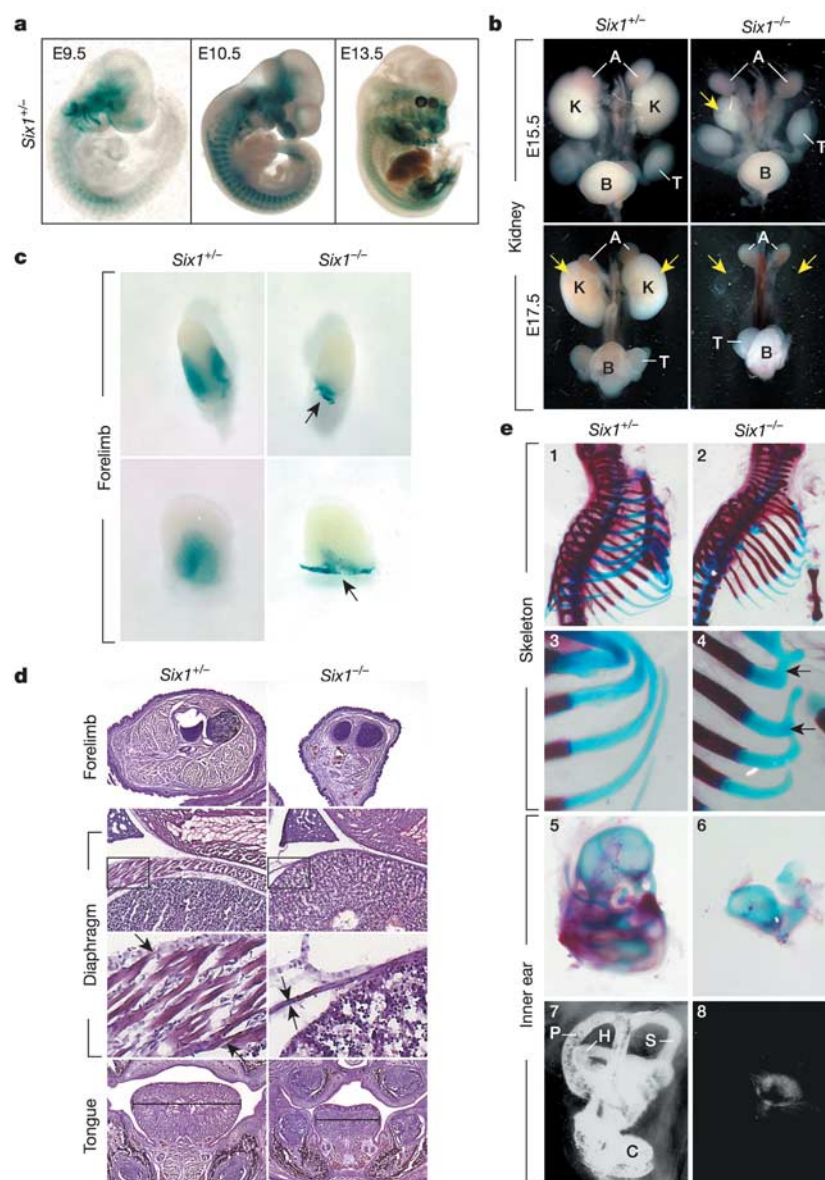


Figure 1 *Six1* is required for the development of multiple organs. **a**, X-gal staining recapitulates endogenous *Six1* expression. **b**, Whole-mount analyses demonstrating either severe reduction or total absence of kidney (K) (arrows). Adrenal (A), bladder (B) and testis (T) appear normal. **c**, X-gal staining of E10.5 embryos reveals absence of muscle precursors in forelimb (arrow; compare left column with right column). **d**, Haematoxylin and eosin histological staining reveals absence of muscle in forelimb, diaphragm and

tongue at e17.5. Boxed areas in the top diaphragm row are magnified in the bottom diaphragm row. **e**, Skeletal staining reveals that the distal ribs are fused (arrows in 4) and inner ear structures are absent. Paint filling demonstrates normal semicircular canals in heterozygotes (panel 7) but not in mutant mice (panel 8). C, cochlea; H, horizontal; P, posterior; S, superior canal.

c), but the expression level of *Gdnf* and *Six2* per cell was reduced, as seen in limb muscle, and *Lbx2* was diminished. Together, our results suggest a proliferation/survival defect during both muscle and kidney development in *Six1*^{-/-} mutant mice. In order to investigate these possibilities, we evaluated the rate of cell proliferation by 5-bromodeoxyuridine (BrdU) labelling *in vivo* at E9.5, E10.5 and E11.5, finding that, in affected tissues, there was a profound decrease (60.5% compared with 21.5%) in cells exhibiting BrdU incorporation (Fig. 3a and data not shown). A TdT-mediated dUTP nick end labelling (TUNEL) assay revealed an increased level of apoptosis in *Six1*^{-/-} mutant mice in regions of hypaxial muscle precursors (fourfold) and developing kidney (3.5-fold) (Fig. 3a), consistent with results for the eye imaginal disc of *Drosophila* with the *so* mutation¹.

To examine further whether these *Six1*-dependent proliferation events were cell autonomous, we used a single-cell nuclear microinjection assay. Nuclear microinjection of purified IgG against *Six1* or *Eya3* into C2C12 cells—a murine myoblast cell line that expresses several members of the *Six/Eya/Dach* family, including *Six1* and *Eya3*—inhibited BrdU incorporation, indicating that both factors were required for effective cell-autonomous proliferation of C2C12 cells (Figs 3b and 4). To confirm these findings, we designed and tested small interfering (si)RNAs against specific murine sequences of *Six1* and *Eya3*. We found that microinjecting the optimal siRNAs directly into the nucleus resulted in consistently efficacious knock-down of RNA transcripts within 24 h of siRNA injection. The targeted transcripts were undetectable by polymerase chain reaction with reverse transcription (RT-PCR) using RNA from cells injected with specific siRNAs, but not cells injected with control siRNAs (Supplementary Fig. 2a). The specificity of the effect was confirmed

by 'rescue' with microinjection of purified bacterially expressed holoproteins (see below, Fig. 4i, and data not shown). Both *Six1* and *Eya3* siRNAs inhibited proliferation of C2C12 cells (Fig. 3c). In contrast, using AIF-1 staining as a marker of apoptosis, anti-*Six1* or anti-*Eya3* IgGs or specific siRNAs did not cause detectable apoptosis of C2C12 cells (data not shown), indicating a major cell-autonomous role of *Six1* and *Eya3* in proliferation regulation in these cells.

Because of the limited material available for the study of early

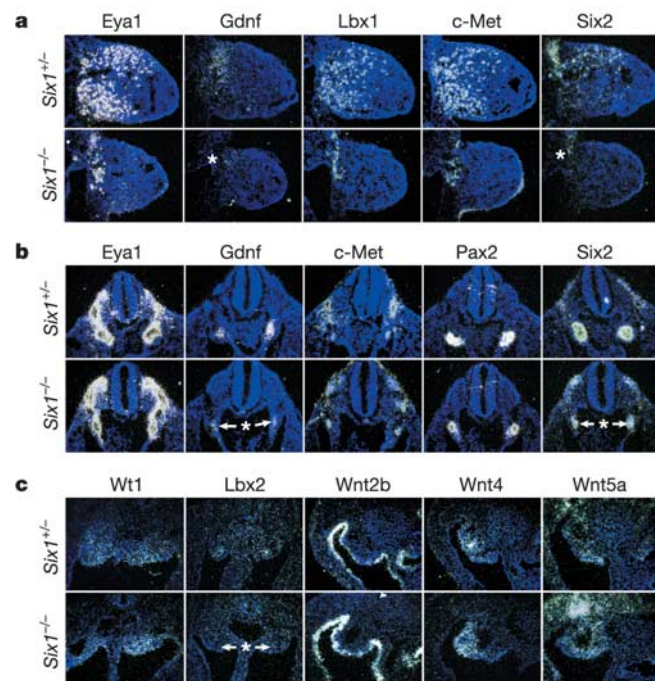


Figure 2 Molecular analyses of the developmental muscle and kidney defects. **a**, *In situ* hybridization at E10.5 reveals reduction of limb-muscle precursor cells, which are ectopically located at the proximal region of limb bud. Expression of *Eya1*, *Lbx1* and *c-Met* is similar in the homozygous mutant, whereas *Gdnf* and *Six2* expression appears reduced (asterisk) compared with heterozygotes. **b**, *In situ* analyses at E10.5 reveals a reduced number of cells but similar expression levels for *Eya1*, *c-Met* and *Pax2*, but reduced expression of *Gdnf* and *Six2* (asterisk) in developing kidney. **c**, *Lbx2* expression is absent in the mutant kidney at E11.0 whereas expression of *Wt1* and different Wnt genes has no apparent difference.

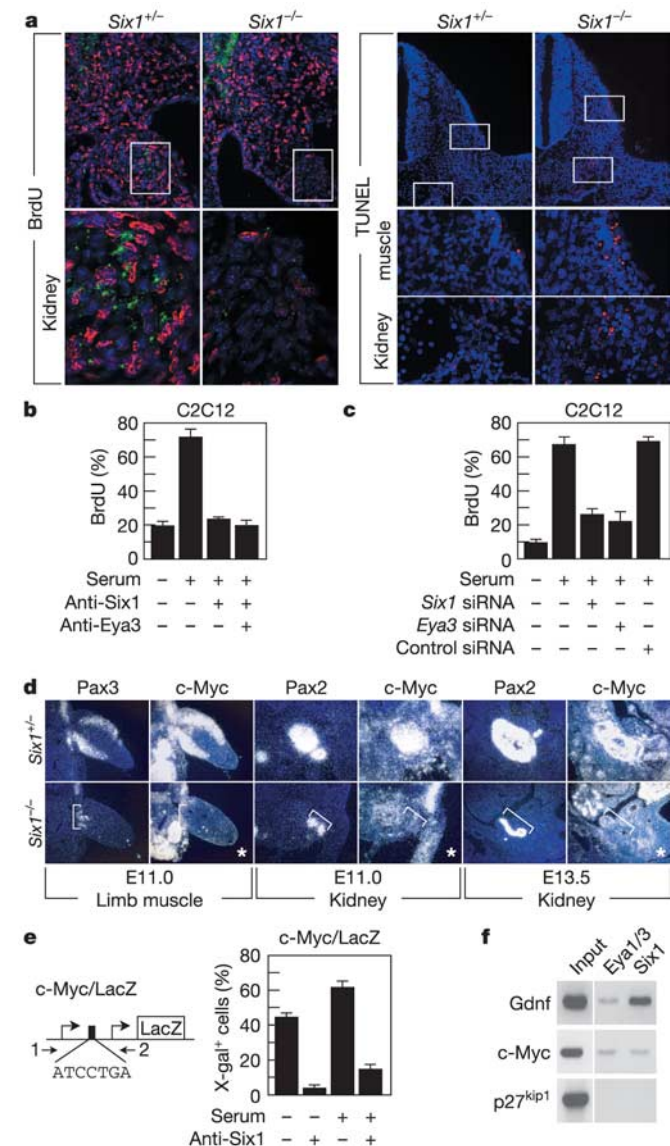


Figure 3 Decreased cell proliferation, increased cell death and diminished *c-Myc* gene expression. **a**, BrdU and TUNEL assays reveal decreased BrdU incorporation, metanephrogenic mesenchyme (red; compare bottom left with bottom right in BrdU panel) and increased apoptosis in both hypaxial muscle precursor cells and metanephrogenic mesenchyme (red; compare left and right kidney muscle column in TUNEL panel). Boxed areas are magnified in the lower panels for BrdU and TUNEL. Blue, DAPI staining; green, X-gal staining. **b**, Microinjection assay using IgG specific for *Six1* and *Eya3* reveals their requirement for proliferation as determined by BrdU incorporation. **c**, Similar results were obtained using siRNAs specific for *Six1* and *Eya3*, but not control siRNA. **d**, *In situ* hybridization showing diminished *c-Myc* expression at E11.0 and E13.5. *Pax3* or *Pax2* on adjacent sections indicate the presence of precursor cells (bracket). **e**, The *c-Myc* promoter contains a *Six1*-binding site and responds to anti-*Six1* IgG. **f**, A ChIP assay reveals recruitment of *Six1* and *Eya1/3* to both *c-Myc* and *Gdnf* regulatory regions in C2C12 cells. The p27^{kip1} coding sequence exhibited no enrichment, and serves as a control.

organogenesis, our initial efforts to determine potential Six1 gene targets used C2C12 cells after transient overexpression of either VP16-Six1 or engrailed repressor domain-Six1 fusion proteins. Transiently transfected cells were sorted based on the co-expression of enhanced green fluorescent protein (EGFP) marker. Gene expression profiles of these cell populations were analysed using Affymetrix Genechips (murine 74Av2). This analysis provided a number of potential targets, of which we evaluated 50 of the most statistically significant candidates, including *Gdnf*, *Six2* and *c-Myc*, by *in situ* hybridization of E11.0 and E13.5 *Six1*^{-/-} mutant and wild-type littermate controls. Expression of *c-Myc* in the mutant was virtually absent in both limb-muscle precursor cells and developing kidney precursor cells (E11.0, E13.5) (Fig. 3d). Expression of *Pax3* and *Pax2* in adjacent sections provided confirmation of the expression of both transcripts in the same muscle and kidney precursor cells that failed to express *c-Myc* (Fig. 3d), consistent with the functions of the *Myc* and *Gdnf* genes during kidney organogenesis^{24,31,32}.

To investigate whether *c-Myc* and *Gdnf* could be direct Six1 target genes, we first analysed the promoter sequences of both genes, finding that the *c-Myc* promoter harbours a Six consensus site (ATCCTGA) immediately 3' to the E2F site (Fig. 3e). Similarly, three perfect consensus Six1 sites were identified in the first intron

of the *Gdnf* gene. To evaluate the potential roles of the *Six1* gene, we used the single-cell nuclear microinjection assay³³⁻³⁵, finding that the expression of both *c-Myc* and *Gdnf* reporters was dependent on the expression of Six1 (Fig. 3e, and data not shown), as injection of anti-Six1 IgG decreased the basal reporter gene expression in C2C12 cells. Furthermore, chromatin immunoprecipitation (ChIP) in C2C12 cells demonstrated that both Six1 and Eya1/3 were present on *c-Myc* and *Gdnf* regulatory sequences (Fig. 3f).

Functions of Eya phosphatase activity

Because *So* and *Eya* synergistically induce eye formation in *Drosophila*, where genetic linkage of *so-eya-dac* is well established^{10,11}, we wished to determine the roles of *Eya* and *Dach* in Six1-mediated proliferation events. Co-transfection analysis using the well-established Six1-responsive myogenin promoter¹⁸ showed that Six1 actively repressed reporter gene expression in heterologous 293 cells, whereas expression of *Eya3* reversed repression and displayed some degree of activation above baseline levels (Fig. 4a), suggesting a functional interaction between Six1 and *Eya3* for downstream gene activation.

In order to study the mechanism of the effects of *Eya3*, we first analysed the primary amino acid sequence of all known members of the *Eya* family. These sequences contained a consensus of two

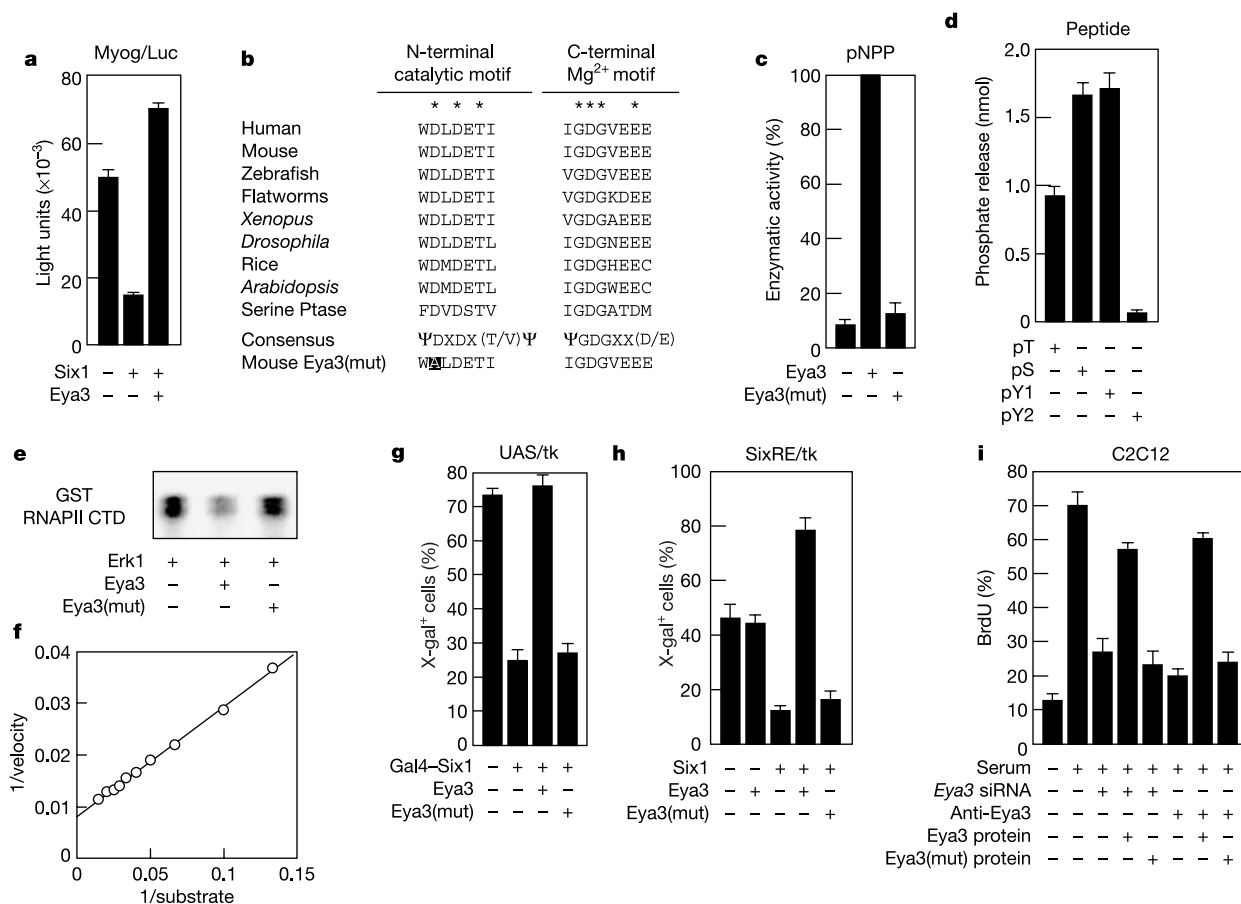


Figure 4 Eya has intrinsic phosphatase activity and this is required for Six1-mediated gene activation and cell proliferation. **a**, Transient transfection assays indicating that *Eya3* releases Six1 repression, causing mild stimulation. **b**, Eya protein sequences share a conserved enzymatic motif and a metal-binding motif identical to phosphoserine phosphatase. **c**, Bacterially expressed wild-type *Eya3* protein, but not the mutant, is able to hydrolyse the phosphatase substrate p-nitrophenylphosphate (pNPP). **d**, *Eya3* exhibits dual specificity *in vitro* using phosphothreonine (pT), phosphoserine (pS) and phosphotyrosine (pY1, not pY2) peptide substrates measured by phosphate release.

e, *Eya3* wild-type protein, but not mutant, dephosphorylates a ³²P-labelled C-terminal repeat of RNA polymerase II (GST-RNAP II CTD) using Erk1/2. **f**, Double-reciprocal plot of *Eya3* phosphatase activity using pNPP substrate (7.5–70 mM). **g**, **h**, *Eya3*(mut) failed to activate the Six1-mediated Gal4-UAS reporter or a reporter dependent on a Six1-response element. **i**, Microinjection of either anti-*Eya3* IgG or *Eya3* siRNA inhibits C2C12 cell proliferation. Co-injection of bacterially expressed wild-type *Eya3* protein, but not the mutant, rescues the effect.

sequence motifs corresponding to the haloacid dehalogenase (HAD) family of phosphohydrolases, which is composed of phosphatases, P-type ATPases, L-2-HADs and epoxide hydrolase, among others^{36,37} (Fig. 4b). The HAD family is characterized by a Ψ DXXX(T/V) Ψ (where Ψ represents a hydrophobic residue) motif near the amino terminus, with the first aspartate acting as a phosphoryl acceptor during substrate dephosphorylation (Fig. 4b). The fourth residue in this motif is commonly an aspartate in phosphatases (Ψ DXDX(T/V) Ψ), a threonine in P-type ATPases (Ψ DKTGT Ψ) and a tyrosine in L-2-HADs³⁷. In addition, phosphatase and ATPases contain a conserved GDGXXD motif near the

carboxy terminus, whereas HADs contain a different SSXXXX sequence. The motifs at the N and C termini of an analogous Eya conserved domain (WDLDETI and GDGVVEE) suggest a phosphatase function, akin to that of phosphatase members of this family, such as phosphoserine phosphatase (PSP), FCP1 and small CTD phosphatases (SCPs)^{37–39}.

PSP catalyses the dephosphorylation of L-phosphoserine in the biosynthetic pathway for L-serine³⁷, whereas SCPs and FCP1 dephosphorylate the RNA polymerase II (RNAP II) C-terminal domain (CTD) during transcriptional regulation^{38–40}. The crystal structure of PSP reveals that the first aspartate (indicated in bold font) in the N-terminal motif (Ψ DXDX(T/V) Ψ) is responsible for nucleophilic attack on the phosphate, whereas the second aspartate (Ψ DXDX(T/V) Ψ) stabilizes the leaving group during the dephosphorylation reaction⁴¹. The aspartates in the C-terminal motif (GDGXXD) facilitate the coordination of the Mg^{2+} ion in the active site. Correspondingly, mutation of one or more of these conserved aspartates in Eya is expected to cause a loss of phosphatase activity, as in the case of PSP and FCP1 (refs 37–39).

To test experimentally this putative enzymatic activity of Eya, we first used the artificial phosphatase substrate p-nitrophenylphosphate (pNPP) (Fig. 4c), and showed that the purified, bacterially expressed Eya1 and Eya3 holoproteins indeed exhibit phosphatase activity, and that a single point mutation of the first aspartate to alanine (Eya3(mut)) in the Ψ DXDX(T/V) Ψ catalytic motif abolishes this activity (Fig. 4c, e, and data not shown). Furthermore, Eya3 phosphatase displayed dual specificity *in vitro*, using phosphothreonine/serine and phosphotyrosine peptide substrates (Fig. 4d). Similar to SCPs and FCP1, Eya1 and Eya3 could dephosphorylate purified RNAP II CTD polypeptide labelled by phosphorylation *in vitro* with ERK 1/2 (Fig. 4e, and data not shown). Kinetic parameters of Eya3 were assessed by steady-state measurements of the rate of pNP production as a function of pNPP concentration (7.5–70 mM) (Fig. 4f). A double reciprocal plot of initial 1/velocity versus 1/[pNPP] fitted well to a straight line ($R^2 = 0.9975$) and yielded K_m (Michaelis constant), K_{cat} (catalytic rate constant) and V_{max} (velocity of enzyme-catalysed reaction at infinite concentration of substrate) values of 25 mM, $0.074 s^{-1}$ and $119 mM s^{-1}$, respectively. Compared with known CTD phosphatases, the turnover number is approximately 37-fold higher than that reported for pNPP hydrolysis by *Saccharomyces cerevisiae* FCP1 (K_{cat} of $0.002 s^{-1}$) and about 27-fold lower than that for *Schizosaccharomyces pombe* FCP1 (K_{cat} of $2 s^{-1}$). The K_m for pNPP binding is comparable to that of *S. pombe* FCP1 (K_m of 19 mM) and about 2.5-fold lower than that of *S. cerevisiae* FCP1 (K_m of 60 mM). In all, the catalytic efficiency of Eya1 (K_{cat}/K_m) on pNPP substrate is in the range of the related members of this specific family of phosphatases, including that of *S. cerevisiae* and *S. pombe* FCP1 (refs 38, 40).

To determine whether this enzymatic activity might be functionally required for the effects of Eya3 on gene transcription, we evaluated the ability of Eya3 to overcome the inhibitory actions of a Gal4–Six1 fusion protein on a UAS-dependent reporter, using the single-cell nuclear microinjection assay. Notably, wild-type Eya3 protein fully reversed the inhibitory actions of Gal4–Six1, whereas Eya3(mut) failed to do so (Fig. 4g). Similarly, Six1 exerted a repressive effect on a reporter driven by Six1-responsive elements, and this effect was fully reversed by the actions of Eya3, but not by Eya3(mut) (Fig. 4h). Immunoprecipitation and nuclear translocation analyses provided evidence that the point mutation (Eya3(mut)) did not disrupt Six1–Eya3 or Eya3–Dach1 interactions (Fig. 5h, i, and data not shown).

These data raised the possibility that Eya phosphatase activity is required to regulate cell proliferation. Using C2C12 cells as a model, we evaluated the effect of the phosphatase activity using anti-Eya3 IgG or Eya3 siRNAs. Both treatments inhibited C2C12 cell proliferation as assessed by BrdU incorporation, and this was entirely

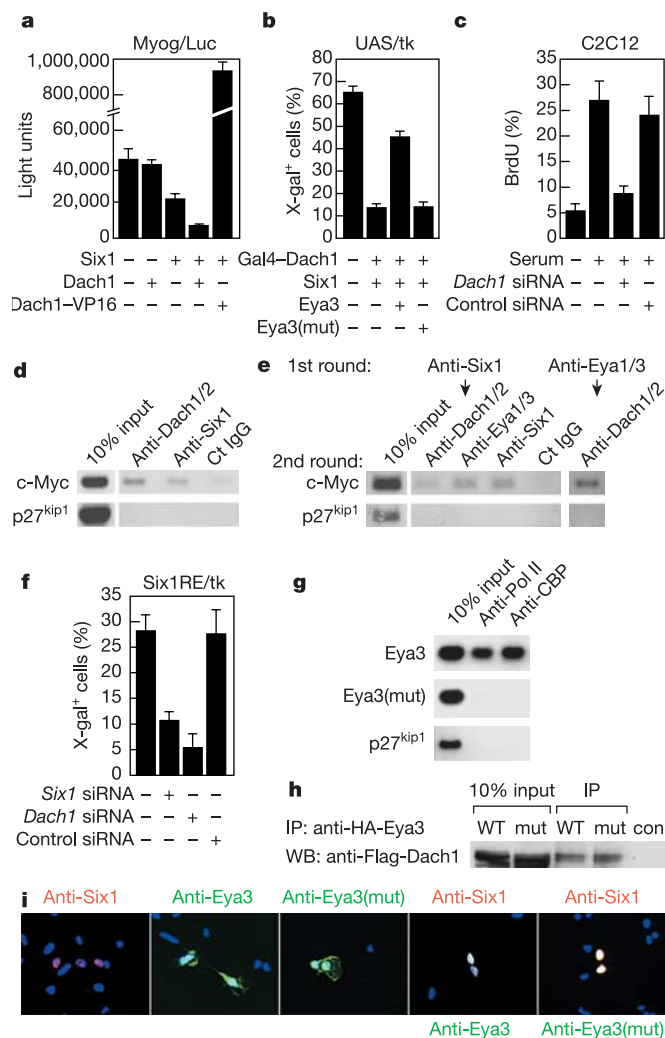


Figure 5 Eya3 phosphatase activity is required to recruit CBP and relieve Dach-mediated repression. **a**, Six1 interacts with Dach1 in a mammalian two-hybrid experiment. **b**, Gal4–Dach repression of a UAS-dependent reporter is reversed by wild-type Eya3 protein, but not Eya3(mut), in microinjection assays. **c**, Microinjection of *Dach1*-specific siRNA inhibits cell proliferation. **d**, Dach1/2 is present on the *c-Myc* promoter along with Six1 using a one-step ChIP assay in C2C12 cells. **e**, Two-step chip assay using either specific anti-Six1 IgG or anti-Eya1/3 IgG, indicating that Dach1/2 is co-recruited along with Six1 and Eya1/3, respectively, onto the *c-Myc* promoter. **f**, Microinjection of *Dach1* siRNA reduced the expression of Six1-dependent reporter expression. **g**, ChIP assay performed after co-transfecting vectors including Six1-dependent reporter, Six1, Dach1 and Eya3 or Eya3(mut) into 293 cells. Co-expression of wild-type Eya3, but not Eya3(mut), recruited the co-activator CBP, activating gene expression as indicated by the presence of RNA polymerase II (Pol II). **h**, Co-immunoprecipitation indicating an interaction of Dach1 with wild-type Eya or mutant protein. **i**, Immunocytochemistry reveals the efficient nuclear translocation of both wild-type and mutant Eya3 protein by co-expression of Six1.

reversed by co-injection of Eya3 wild-type protein; however, the Eya3(mut) protein failed to rescue the block to proliferation (Fig. 4i). These data suggest that Eya phosphatase activity is essential for regulating precursor cell proliferation.

A mammalian two-hybrid assay using the myogenin Six1-regulated reporter revealed that although Dach1 co-expression enhanced Six1-dependent repression, VP16–Dach1 acted as a strong activator, confirming a functional interaction between Six1 and Dach1 (Fig. 5a). To investigate the potential role of Eya in Six–Dach-mediated repression, we used single-cell nuclear microinjection experiments, finding that a Gal4–Dach fusion protein,

together with Six1, strongly repressed UAS-tk reporter gene expression. Co-expression of wild-type Eya3 protein reversed this effect (Fig. 5b). In contrast, Eya3(mut) failed to prevent the repression activity (Fig. 5b), suggesting that Eya phosphatase function is required to reverse Dach–Six-mediated repression.

We therefore sought to evaluate a potential functional role of Dach1 in C2C12 cells. Nuclear microinjection of anti-Dach1 IgG or *Dach1* siRNA revealed that *Dach1* was required for effective serum-stimulated proliferation (Fig. 5c, see also Supplementary Fig. 2a and data not shown). This raised the possibility that both Dach and Eya are required for Six1-mediated gene activation. Therefore, we first used ChIP assays to examine whether both Dach and Eya were present on the *Six* target genes in C2C12 cells. Using IgGs that recognized both Eya1 and Eya3 or Dach1 and Dach2, we found that Six1 along with Eya1/3 and Dach1/2 were present on both *c-Myc* and *Gdnf* regulatory regions (Fig. 5d, and data not shown). To ascertain co-recruitment, a sequential ChIP analysis was performed³⁵, revealing co-occupancy of Six1 with Dach1/2 and Eya3, as well as co-occupancy of Eya1/3 and Dach1/2, on both *c-Myc* and *Gdnf* regulatory regions (Fig. 5e, and data not shown). We then tested whether Dach has a role in gene activation using single-cell nuclear microinjection assays in C2C12 cells. We found that a specific *Dach1* siRNA inhibited reporter gene expression driven by Six1-responsive elements comparable to that observed with anti-Six1 IgG (Fig. 5f), indicating that *Dach1* is required for activation of Six1 target genes. These results are consistent with the finding that Dach and Eya can interact with CREB-binding protein (CBP) and synergistically enhance Gal4–Six5 function on a UAS reporter in a transient transfection assay⁴².

The requirement of the *Dach1* co-repressor to activate transcription prompted the question of whether Eya phosphatase activity is required to modulate the recruitment of other co-repressors or co-activators to Six target genes. To begin to investigate this issue, we transfected a reporter that was dependent on a Six1-response element into 293 cells, along with expression vectors encoding Six1, Dach1 and either wild-type Eya3 or Eya3(mut) (Supplementary Fig. 2b), and used ChIP to evaluate the presence of co-repressors and co-activators. In cells overexpressing wild-type Eya3, the CBP co-activator was recruited along with gene activation, as indicated by recruitment of polymerase II (Fig. 5g). In contrast, Eya3(mut), although expressed at a comparable level, failed to permit recruitment of CBP and polymerase II (Fig. 5g, see also Supplementary Fig. 2b). Therefore, the phosphatase activity of Eya is required to switch Six1 function from repression to activation, permitting recruitment of co-activators, including CBP.

Six1 and Eya1 synergistically regulate organogenesis

We wished to assess further the Eya–Six interactions in a genetic model. Because *Eya1*^{−/−} mice represent the best-characterized mutant model³⁰, the mutant phenotype of which clearly resembles that of *Six1*^{−/−} mice in renal and otic development, we elected to cross *Six1*^{+/−} and *Eya1*^{+/−} mice. *Eya1*^{+/−} *Six1*^{+/−} double heterozygous mice were crossed to obtain *Eya1*^{−/−} *Six1*^{−/−} mice. We found that the double heterozygous mice have a defect in kidney development (Fig. 6b), which is not observed in single heterozygotes for either gene deletion alone, suggesting that Six1 and Eya1 act in the same genetic pathway. Notably, there is a complete absence of all hypaxial muscle in *Six1*^{−/−} *Eya1*^{−/−} double-gene-deleted mice and severe reduction of epaxial muscle (Fig. 6a, and data not shown), a phenotype resembling that of *Myog*^{−/−}, *MyoD*^{−/−} *Myf5*^{−/−} and *Pax3*^{−/−} *Myf5*^{−/−} mutants^{43–46}. Intriguingly, although mutation of *Six1* or *Eya1* has minimal or no effect on pituitary development, mice with both genes deleted have a pituitary that is approximately 5–10-fold smaller by volume than the wild-type gland (Fig. 6a), an indication that Six1 and Eya1 also function in distinct pathways. Together these data provide genetic evidence of Six–Eya interactions during vertebrate organogenesis.

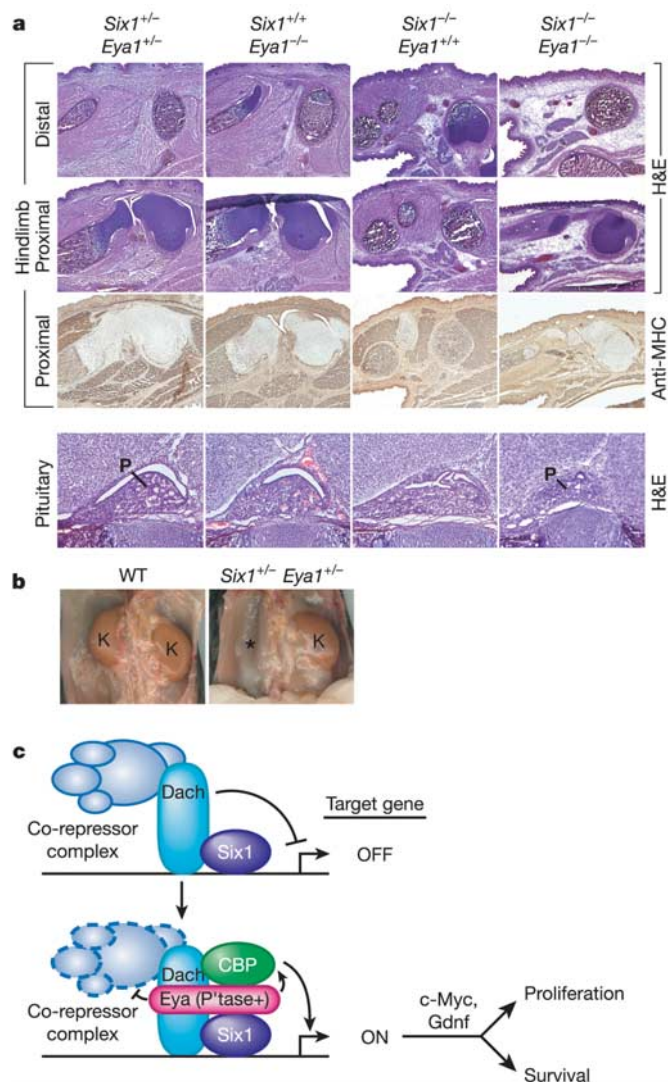


Figure 6 Genetic interaction between *Six1* and *Eya1* in regulating muscle, pituitary and kidney development. **a**, Haematoxylin and eosin, and anti-myosin heavy chain (anti-MHC) staining of E17.5 embryos reveals a dose-dependent effect in muscle and pituitary formation. The *Eya1*^{−/−} single mutant has no detectable muscle phenotype whereas the *Six1*^{−/−} single mutant has reduced muscle mass in the hindlimb. Double mutants (*Six1*^{−/−} *Eya1*^{−/−}) have no detectable limb muscle. Similarly, neither *Six1*^{−/−} nor *Eya1*^{−/−} mutants exhibit severe pituitary (P) defects; however, double mutants exhibit a marked decrease in cell population. **b**, A compound *Six1* and *Eya1* heterozygote reveals kidney (K) hypoplasia, often with unilateral ablation (asterisk). **c**, Model of Six–Eya–Dach-regulated target gene activation. In the absence of functional Eya, Dach recruits a co-repressor complex and inhibits target gene expression. Recruitment of Eya phosphatase co-recruits other co-activators, including CBP, permitting expression of target genes essential for precursor cell proliferation and survival.

Discussion

In addition to the large repertoire of tissue-specific, DNA-binding transcription factors, a large, intricately regulated network of co-activator and co-repressor complexes combinatorially modulate transcription in a tissue- and promoter-specific fashion^{47–49}. The well-defined genetic relationship between *Six–Eya–Dach* has been exploited here, revealing a novel aspect of this co-regulatory network. We have demonstrated that the ability of *Six1* to act both as a repressor or activator is based, at least in part, on the recruitment of opposing cofactors; that is, the *Dach* co-repressors and the *Eya* co-activators. *Eya*, however, can act to reverse *Dach* co-repressor function on the basis of its intrinsic phosphatase activity, allowing it to function as a required co-activator and to permit the recruitment of other co-activators, including CBP. The *Six–Eya–Dach* regulatory network thereby defines a molecular mechanism by which a recruited co-activator with phosphatase function can serve to derepress target genes (Fig. 6c). Thus, as well as the large number of cellular processes regulated by diverse protein phosphatases⁵⁰, a phosphatase can also serve as a promoter-specific transcriptional co-activator. Our results provide a conceptual framework for understanding the *Six–Eya–Dach* genetic interactions that are likely to be prototypic of other co-activator/co-repressor strategies in mammalian organogenesis. □

Methods

Generation of *Six1*^{−/−} mice

The procedure for generating *Six1*^{−/−} mice was essentially the same as described previously¹⁴. Briefly, targeting construct was generated by replacing partial *Six* domain and homeodomain with IRES-LacZ/PGK-neo sequence of *Six1* genomic locus from a 129sv genetic background. Five independent R1 embryonic stem clones (a gift of J. D. Marth), of which three were used for generating knockout mice, were identified by Southern blot, screening with both 5′ and 3′ outside probes. All three lines showed identical phenotypes.

Antibodies and experimental procedures

Immunohistochemistry, *in situ* hybridization, skeletal preparation and paint filling were carried out essentially as described¹⁴. The anti-Eya1, anti-Six1 (Santa Cruz), and anti-Eya3 (generated by immunizing rabbits with a synthetic peptide (KDADDQARKNMTVKNRGK)) antibodies were specific on western blot analysis. All antibodies were purified for both single-cell nuclear microinjection assay and ChIP assay.

BrdU and TUNEL assays

BrdU and TUNEL assays were performed as described previously¹⁴. Briefly, pregnant female mice were injected with 0.1 mg g^{−1} body weight of BrdU/PBS. Embryos were isolated and fixed in 10% neutral buffered formalin overnight. Ten-micrometre cryostat sections were used for both BrdU staining and TUNEL assays. All sections were counter stained with 4,6-diamidino-2-phenylindole (DAPI) for cell counting or photography.

Single-cell nuclear microinjection and ChIP assays

The single-cell nuclear microinjection and ChIP assays were performed as described with >300 cells injected for each point¹⁴. *Six1*, *Eya3* and *Dach1* siRNAs correspond to DNA sequences 5′-AAGAACGAGAGCGTCTCAAG-3′, 5′-AACAGTGATGCTGAGACCACA-3′ and 5′-AAAGTGGCTTCTTTACGGTG-3′, respectively. Nonspecific siRNA is from Dharmacon. PCR primers for *c-Myc* promoter and control primers corresponding to the *p27*^{Kip1} coding region were as described previously¹⁴. Primers used for *Six1*-dependent reporters were 5′-CTTTATGTTTGTGGCGTCTCCAT-3′ and 5′-AATGTATCTTAATGGTACTGTAAC-3′. The negative control primers corresponded to 2 kilobases downstream of the *Six1* elements, and the sequences were 5′-AGCCATACCACATTTGTAGAGG-3′ and 5′-GACGATAGTCATGCCCGCG-3′. *Gdnf* primers were 5′-TTGTGACTCTTGAGAA GGGTG-3′ and 5′-TAGCCAGAGCAATTCGACAAC-3′. Proliferation assay after single-cell injection was done by measuring BrdU immunoactivity as described previously¹⁴.

In vitro phosphatase assays

In vitro phosphatase assays were the same as described previously³⁹. The *Eya3* phosphatase kinetic parameters were estimated using pNPP substrate ranging from 7.5 to 70 mM⁴⁰. Briefly, the reaction mixture containing 2.5 μg purified *Eya3* protein in 200 μl phosphatase buffer (50 mM Tris-HCl, pH 5.3, 10 mM MgCl₂, 10% glycerol, 0.5 mM dithiothreitol) and pNPP substrate was incubated for 60 min at 37°C. The reaction was stopped by adding 800 μl of 0.2 M NaOH and 50 mM EDTA, and the pNP production was determined by measuring A410 and standard. Phosphate release from the phosphothreonine (pT; KRpTIRR), phosphoserine (pS; RRApSVA), or phosphotyrosine peptides (pY1; RRLIEDAEPYAARG, pY2; TSTEPQYQPGENL; Upstate Biotech) was assayed as described³⁹ according to the manufacturer's suggestion.

Received 19 May; accepted 16 September 2003; doi:10.1038/nature02083.

1. Cheyette, B. N. *et al.* The *Drosophila* sine oculis locus encodes a homeodomain-containing protein

required for the development of the entire visual system. *Neuron* **12**, 977–996 (1994).

- Bonini, N. M., Leiserson, W. M. & Benzer, S. The eyes absent gene: genetic control of cell survival and differentiation in the developing *Drosophila* eye. *Cell* **72**, 379–395 (1993).
- Mardon, G., Solomon, N. M. & Rubin, G. M. dachshund encodes a nuclear protein required for normal eye and leg development in *Drosophila*. *Development* **120**, 3473–3486 (1994).
- Bonini, N. M., Bui, Q. T., Gray-Board, G. L. & Warrick, J. M. The *Drosophila* eyes absent gene directs ectopic eye formation in a pathway conserved between flies and vertebrates. *Development* **124**, 4819–4826 (1997).
- Shen, W. & Mardon, G. Ectopic eye development in *Drosophila* induced by directed dachshund expression. *Development* **124**, 45–52 (1997).
- Oliver, G. *et al.* *Six3*, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. *Development* **121**, 4045–4055 (1995).
- Xu, P. X., Woo, I., Her, H., Beier, D. R. & Maas, R. L. Mouse *Eya* homologues of the *Drosophila* eyes absent gene require Pax6 for expression in lens and nasal placode. *Development* **124**, 219–231 (1997).
- Hammond, K. L., Hanson, I. M., Brown, A. G., Lettice, L. A. & Hill, R. E. Mammalian and *Drosophila* dachshund genes are related to the *Ski* proto-oncogene and are expressed in eye and limb. *Mech. Dev.* **74**, 121–131 (1998).
- Heanue, T. A. *et al.* Synergistic regulation of vertebrate muscle development by *Dach2*, *Eya2*, and *Six1*, homologs of genes required for *Drosophila* eye formation. *Genes Dev.* **13**, 3231–3243 (1999).
- Chen, R., Amoui, M., Zhang, Z. & Mardon, G. Dachshund and eyes absent proteins form a complex and function synergistically to induce ectopic eye development in *Drosophila*. *Cell* **91**, 893–903 (1997).
- Pignoni, F. *et al.* The eye-specification proteins *So* and *Eya* form a complex and regulate multiple steps in *Drosophila* eye development. *Cell* **91**, 881–891 (1997).
- Ohto, H. *et al.* Cooperation of *six* and *eya* in activation of their target genes through nuclear translocation of *Eya*. *Mol. Cell. Biol.* **19**, 6815–6824 (1999).
- Davis, R. J., Shen, W., Heanue, T. A. & Mardon, G. Mouse *Dach*, a homologue of *Drosophila* dachshund, is expressed in the developing retina, brain and limbs. *Dev. Genes Evol.* **209**, 526–536 (1999).
- Li, X., Perissi, V., Liu, F., Rose, D. W. & Rosenfeld, M. G. Tissue-specific regulation of retinal and pituitary precursor cell proliferation. *Science* **297**, 1180–1183 (2002).
- Zhu, C. C. *et al.* *Six3*-mediated auto repression and eye development requires its interaction with members of the Groucho-related family of co-repressors. *Development* **129**, 2835–2849 (2002).
- Lagutin, O. V. *et al.* *Six3* repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. *Genes Dev.* **17**, 368–379 (2003).
- Pohl, M., Stuart, R. O., Sakurai, H. & Nigam, S. K. Branching morphogenesis during kidney development. *Annu. Rev. Physiol.* **62**, 595–620 (2000).
- Laclef, C. *et al.* Altered myogenesis in *Six1*-deficient mice. *Development* **130**, 2239–2252 (2003).
- Daston, G., Lamar, E., Olivier, M. & Goulding, M. Pax-3 is necessary for migration but not differentiation of limb muscle precursors in the mouse. *Development* **122**, 1017–1027 (1996).
- Schafer, K. & Braun, T. Early specification of limb muscle precursor cells by the homeobox gene *Lbx1*. *Nature Genet.* **23**, 213–216 (1999).
- Gross, M. K. *et al.* *Lbx1* is required for muscle precursor migration along a lateral pathway into the limb. *Development* **127**, 413–424 (2000).
- Brohmann, H., Jagla, K. & Birchmeier, C. The role of *Lbx1* in migration of muscle precursor cells. *Development* **127**, 437–445 (2000).
- Dietrich, S. *et al.* The role of SF/HGF and c-Met in the development of skeletal muscle. *Development* **126**, 1621–1629 (1999).
- Pichel, J. G. *et al.* Defects in enteric innervation and kidney development in mice lacking *GDNF*. *Nature* **382**, 73–76 (1996).
- Xu, P. X. *et al.* *Eya1*-deficient mice lack ears and kidneys and show abnormal apoptosis of organ primordia. *Nature Genet.* **23**, 113–117 (1999).
- Torres, M., Gomez-Pardo, E., Dressler, G. R. & Gruss, P. Pax-2 controls multiple steps of urogenital development. *Development* **121**, 4057–4065 (1995).
- Schuchardt, A., D'Agati, V., Larsson-Blomberg, L., Costantini, F. & Pachnis, V. Defects in the kidney and enteric nervous system of mice lacking the tyrosine kinase receptor *Ret*. *Nature* **367**, 380–383 (1994).
- Kreidberg, J. A. *et al.* *WT-1* is required for early kidney development. *Cell* **74**, 679–691 (1993).
- Kispert, A., Vainio, S. & McMahon, A. P. Wnt-4 is a mesenchymal signal for epithelial transformation of metanephric mesenchyme in the developing kidney. *Development* **125**, 4225–4234 (1998).
- Davies, J. A. & Bard, J. B. The development of the kidney. *Curr. Top. Dev. Biol.* **39**, 245–301 (1998).
- Stanton, B. R., Perkins, A. S., Tassarollo, L., Sassoon, D. A. & Parada, L. F. Loss of N-myc function results in embryonic lethality and failure of the epithelial component of the embryo to develop. *Genes Dev.* **6**, 2235–2247 (1992).
- Davis, A. C., Wims, M., Spotts, G. D., Hann, S. R. & Bradley, A. A null c-myc mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice. *Genes Dev.* **7**, 671–682 (1993).
- Jepsen, K. *et al.* Combinatorial roles of the nuclear receptor corepressor in transcription and development. *Cell* **102**, 753–763 (2000).
- Kioussi, C. *et al.* Identification of a Wnt/Dvl/β-Catenin → Pitx2 pathway mediating cell-type-specific proliferation during development. *Cell* **111**, 673–685 (2002).
- Baek, S. H. *et al.* Exchange of N-CoR corepressor and Tip60 coactivator complexes links gene expression by NF-κB and β-amyloid precursor protein. *Cell* **110**, 55–67 (2002).
- Collet, J. F., Stroobant, V., Pirard, M., Delpierre, G. & Van Schaftingen, E. A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif. *J. Biol. Chem.* **273**, 14107–14112 (1998).
- Collet, J. F., Stroobant, V. & Van Schaftingen, E. Mechanistic studies of phosphoserine phosphatase, an enzyme related to P-type ATPases. *J. Biol. Chem.* **274**, 33985–33990 (1999).
- Kobor, M. S. *et al.* An unusual eukaryotic protein phosphatase required for RNA polymerase II and CTD dephosphorylation in *S. cerevisiae*. *Mol. Cell* **4**, 55–62 (1999).
- Yeo, M., Lin, P. S., Dahmus, M. E. & Gill, G. N. A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5. *J. Biol. Chem.* **278**, 26078–26085 (2003).
- Hausmann, S. & Shuman, S. Characterization of the CTD phosphatase Fcp1 from fission yeast. Preferential dephosphorylation of serine 2 versus serine 5. *J. Biol. Chem.* **277**, 21213–21220 (2002).
- Wang, W. *et al.* Structural characterization of the reaction pathway in phosphoserine phosphatase:

- crystallographic “snapshots” of intermediate states. *J. Mol. Biol.* **319**, 421–431 (2002).
42. Ikeda, K., Watanabe, Y., Ohno, H. & Kawakami, K. Molecular interaction and synergistic activation of a promoter by Six, Eya, and Dach proteins mediated through CREB binding protein. *Mol. Cell. Biol.* **22**, 6759–6766 (2002).
 43. Hasty, P. *et al.* Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. *Nature* **364**, 501–506 (1993).
 44. Nabeshima, Y. *et al.* Myogenin gene disruption results in perinatal lethality because of severe muscle defect. *Nature* **364**, 532–535 (1993).
 45. Rudnicki, M. A. *et al.* MyoD or Myf-5 is required for the formation of skeletal muscle. *Cell* **75**, 1351–1359 (1993).
 46. Tajbakhsh, S., Rocancourt, D., Cossu, G. & Buckingham, M. Redefining the genetic hierarchies controlling skeletal myogenesis: Pax-3 and Myf-5 act upstream of MyoD. *Cell* **89**, 127–138 (1997).
 47. McKinsey, T. A., Zhang, C. L. & Olson, E. N. Control of muscle development by dueling HATs and HDACs. *Curr. Opin. Genet. Dev.* **11**, 497–504 (2001).
 48. McKenna, N. J. & O’Malley, B. W. Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell* **108**, 465–474 (2002).
 49. Glass, C. K. & Rosenfeld, M. G. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* **14**, 121–141 (2000).

50. Denu, J. M., Stuckey, J. A., Saper, M. A. & Dixon, J. E. Form and function in protein dephosphorylation. *Cell* **87**, 361–364 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank G. N. Gill, J. E. Dixon, M. Yeo, L. Erkman, V. Perissi, L. Olson, F. Liu and V. Kumar for reagents, critical advice and suggestions; P. Myer and M. Fisher for figure and manuscript preparation; C. Nelson for tissue culture; P. Kotol for technical assistance; and H. Taylor for animal care. We are grateful to those who provided us with essential reagents for this study. This work was supported by grants from the NIH to X.L., M.G.R., D.W.R., R.M., A.K.A. and S.K.N. M.G.R. is an investigator with the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.L. (seli@ucsd.edu) or M.G.R. (mrosenfeld@ucsd.edu).

Regulation of phyllotaxis by polar auxin transport

Didier Reinhardt^{1*}, Eva-Rachele Pesce^{1*}, Pia Stieger¹, Therese Mandel¹, Kurt Baltensperger², Malcolm Bennett³, Jan Traas⁴, Jiří Friml⁵ & Cris Kuhlemeier¹

¹Institute of Plant Sciences, University of Bern, Altenbergrain 21, 3013 Bern, Switzerland

²Institute of Pharmacology, University of Bern, Friedbühlstrasse 49, 3010 Bern, Switzerland

³School of Biosciences, University of Nottingham, Nottingham NG7 2RD, UK

⁴Laboratoire de Biologie Cellulaire, INRA, Route de Saint Cyr, 78026 Versailles cedex, France

⁵Zentrum für Molekularbiologie der Pflanzen, Universität Tübingen, Auf der Morgenstelle 3, D-72076 Tübingen, Germany

* These authors contributed equally to this work

The regular arrangement of leaves around a plant's stem, called phyllotaxis, has for centuries attracted the attention of philosophers, mathematicians and natural scientists; however, to date, studies of phyllotaxis have been largely theoretical. Leaves and flowers are formed from the shoot apical meristem, triggered by the plant hormone auxin. Auxin is transported through plant tissues by specific cellular influx and efflux carrier proteins. Here we show that proteins involved in auxin transport regulate phyllotaxis. Our data indicate that auxin is transported upwards into the meristem through the epidermis and the outermost meristem cell layer. Existing leaf primordia act as sinks, redistributing auxin and creating its heterogeneous distribution in the meristem. Auxin accumulation occurs only at certain minimal distances from existing primordia, defining the position of future primordia. This model for phyllotaxis accounts for its reiterative nature, as well as its regularity and stability.

A major determinant of plant architecture is the arrangement of leaves and flowers around the stem, known as phyllotaxis¹. Leaves and flowers are formed by the shoot apical meristem in characteristic patterns such as alternate (distichous), opposite (decussate), or—most frequently—spiral^{1,2}. Spiral phyllotaxis has received much attention from theoreticians because the divergence angle between successive leaves approaches the golden ratio of 137.5°, and the spiral arrangements are characterized by the Fibonacci numbers^{1,3,4}. Models of phyllotaxis explain the regular spacing of leaf primordia with mechanisms based on “the first available space”⁵, “biophysical forces” within the meristem⁶, or “inhibitors of organogenesis” released from young leaf primordia^{3,7}. Although mathematical modelling of phyllotactic mechanisms based on biophysical or biochemical regulation can recreate phyllotactic patterns^{3,6,8}, experimental support for such phyllotactic mechanisms is lacking.

Auxin has been implicated in various aspects of plant development, including embryogenesis^{9–13}, root development^{14,15}, and vascular differentiation¹⁶. Recently, auxin has been shown to induce leaf and flower formation at the shoot apical meristem (SAM)¹⁷. As a trigger of organ initiation, auxin could play a role in determining organ position and thereby phyllotaxis¹⁸.

An instructive role of a signal molecule implies regulated distribution in the tissue. This could be achieved by defined sources of the signal, by regulated degradation, or by controlled transport of the signal to certain sites in the tissue. Auxin has its own polar transport route consisting of specific influx and efflux carriers in the plasma membrane¹⁹. Net auxin flux results from coordinated polar localization of the auxin efflux carrier in the auxin-conducting cells²⁰. This routing of auxin is responsible for basipetal auxin transport from the young leaves towards the root. In the root tip, active polar transport generates an auxin maximum in the columella initials, which regulates meristem development and cell differentiation^{14,15}. Furthermore, auxin efflux carriers mediate gravitropism by creating asymmetric auxin distribution, thereby causing anisotropic growth²¹.

We have previously proposed that active auxin transport in the shoot apex could generate patterned auxin distribution, thereby determining the pattern of organ formation. To allow

valid predictions about the transport and distribution of auxin within the shoot apex, we studied the expression pattern and subcellular localization of the putative *Arabidopsis* influx and efflux carriers, AUXIN RESISTANT1 (AUX1)²² and PIN-FORMED1 (PIN1)²³. In addition, we performed micro-application of auxin to *Arabidopsis* meristems to address the role of auxin in phyllotaxis.

PIN1 expression and localization in *Arabidopsis* apices

In vegetative apices, PIN1 protein was detected in the epidermis and the vasculature of developing leaf primordia, as well as in the meristem L₁ layer and, at lower levels, in L₂ (Fig. 1a). A marked induction at the meristem flank (Fig. 1a) corresponded with the site of incipient and young leaf primordia (Fig. 1b). PIN1 signal in the L₁ and L₂ layers was consistently localized to the apical side of the cells, that is, pointing to the meristem summit (Fig. 1c, arrowheads), except for the cells close to the site of incipient leaf initiation (I₁). In these cells, the signal pointed to I₁ (Fig. 1c, arrows). At the summit itself, the signal was more diffuse than at the flank (Fig. 1c). In the epidermis of young leaf primordia, PIN1 signal was restricted to the apical side of the cells (Fig. 1d, e), as in cotyledon primordia²⁴. At the site of incipient organ formation, PIN1 protein in the inner cell layers was first localized with no obvious direction of polarization (Fig. 1f), but at the time of primordium outgrowth, PIN1 was detected at the side of the cells that points to the centre of the leaf primordia (Fig. 1g). This localization suggests accumulation of auxin in the centre of the young primordia.

During primordium outgrowth, PIN1 became restricted to a narrow file of cells that comprised 2–3 cells in width, corresponding presumably to the provascular strand, and the PIN1 signal in these cells became gradually localized to the basal end of the cells (Fig. 1h). The gradual restriction to narrow cell files and the subcellular polarization of PIN1, first towards the centre and later towards the base of the leaf primordia, is suggestive of a canalization mechanism²⁵. The canalization theory proposes that small local differences in auxin concentration are amplified by a self-reinforcing accumulation mechanism, resulting in local auxin elevation and auxin depletion in the surrounding tissue. As a

consequence, the cells that experience high auxin levels are restricted to narrow stripes that are determined to become auxin-conducting cell files²⁵.

In inflorescence apices, PIN1 expression followed a similar phyllotactic pattern as in vegetative meristems (Fig. 1i, compare with Fig. 1b). Notably, PIN1 induction was already clearly visible in incipient flower primordia at the I_3 stage, at least one plastochron earlier than the induction of the organ marker gene *LEAFY* (*LFY*)²⁶ (Fig. 1j). PIN1 was detected in the vasculature of the inflorescence stem as well as in young flower primordia (Fig. 1k). In the inflorescence meristem, the PIN1 signal was strongest in the L_1 layer, and lower in the interior layers with upregulation at the site of incipient organ formation (arrowhead in Fig. 1k). In young flower primordia, PIN1 expression was elevated at the flank, at the site of sepal induction (Fig. 1l), and later where the inner whorl organs were formed (Fig. 1m).

Organ formation in mutants with auxin-related defects

The similar expression patterns of PIN1 in vegetative and floral apices suggest a comparable function of PIN1 and auxin in the induction of leaves and flowers. However, *pin1* mutants can form leaves, although their size, shape, and position are aberrant^{27,28}. To address the role of auxin in leaf formation, we treated young *pin1* plants with indole-3-acetic acid (IAA), resulting in the formation of normal leaves (Fig. 2a). In addition, we crossed *pin1* mutants with *lfy* mutants, which form leaves and axillary inflorescences instead of flowers²⁶. Treatment of pin-shaped apices of *pin1;lfy* double mutants with IAA resulted in the formation of leaves instead of flowers (Fig. 2b). These leaves had abundant trichomes on their abaxial side (Fig. 2b), and formed pin-shaped lateral branches in their axils (not shown), indicating that they corresponded to cauline leaves. Hence, IAA can trigger the initiation of rosette leaves, cauline leaves, and flowers¹⁷.

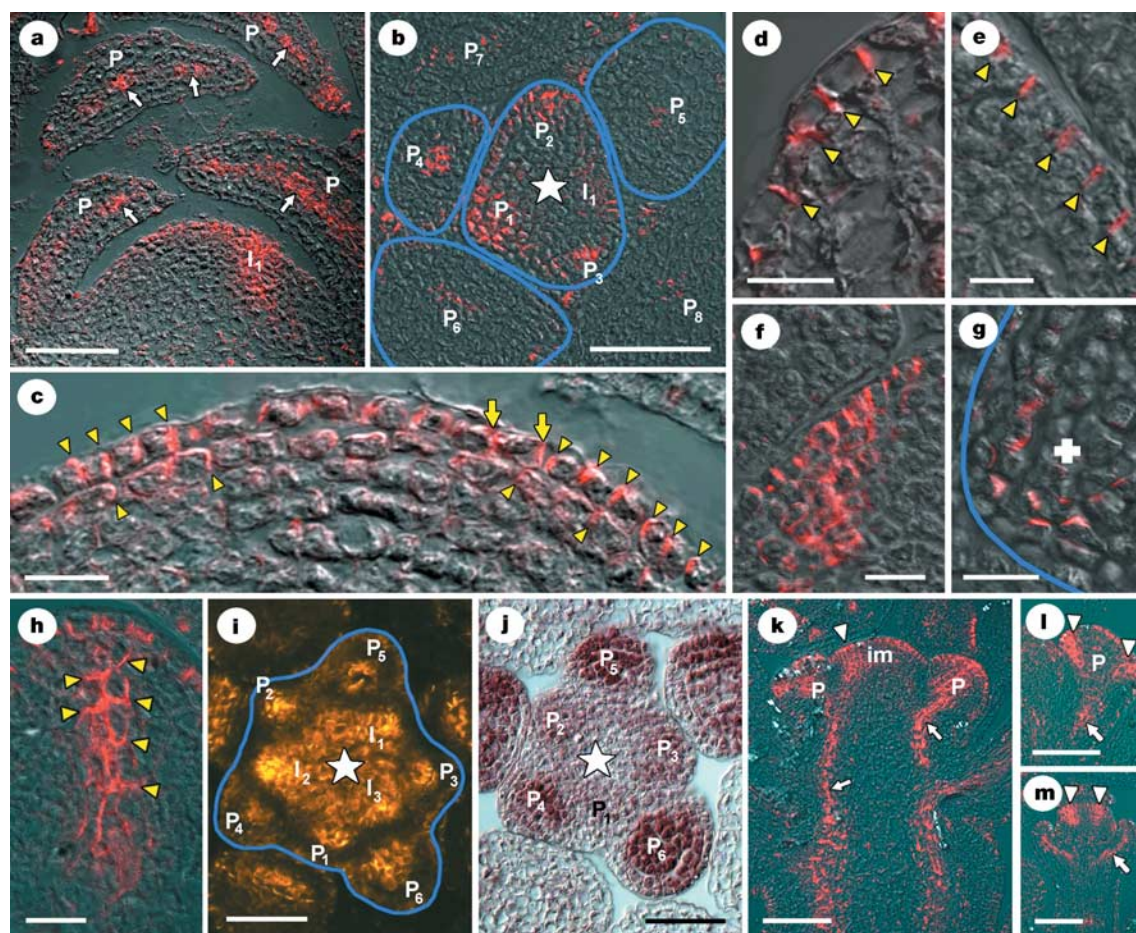


Figure 1 Localization of the PIN1 protein in *Arabidopsis* apices. **a**, Median longitudinal section through a vegetative apex. Note PIN1 signal (red) in the vasculature of leaf primordia and at the site of incipient organ formation (I_1). **b**, Cross-section of an apex as in **a**. The outline of primordia (blue) is highlighted for clarity. Note PIN1 expression in young primordia (P_1 , P_2 and P_3) and the developing central vein of older primordia (P_4 – P_8). **c**, Near-median longitudinal section of the same meristem as depicted in **a**. Yellow arrows indicate discontinuity of PIN1 localization at I_1 . **d**, Longitudinal section through a young leaf primordium showing PIN1 signal on the distal end of cells in the abaxial epidermis. **e**, Tangential longitudinal section through a vegetative apex showing PIN1 signal in the lateral margin of a young leaf primordium. Leaf tip at the top of the image. **f**, Close-up at I_1 of a section as in **a**. **g**, Close-up of the same section as depicted in **b**, corresponding to the P_1 position, imaged with reduced pinhole size. The centre of the leaf primordium is denoted by a cross. **h**, Tangential longitudinal section of a leaf primordium showing localization of PIN1 in the central provascular strand. **i**, Cross-section through an

inflorescence apex with PIN1 signal in the centres of flower primordia (P_1 – P_8), and at the site of incipient flower formation (I_1 – I_3). **j**, Section as in **i** hybridized with an antisense probe against *LEAFY* mRNA. **k**, Longitudinal section through an inflorescence apex. Note PIN1 signal in the vasculature, in the L_1 layer of inflorescence and flower meristems, and at the site of incipient organ formation (arrowhead). **l**, Longitudinal section through an early flower primordium. Signals coincide with the position of incipient sepals (arrowheads) and the developing vascular strand. **m**, Longitudinal section through a later flower primordium. Signal coincides with the position of incipient anthers (arrowheads) and with the central vascular strand of sepals. P_1 – P_8 , leaf primordia in **b** and flower primordia in **i** and **j**, with P_1 being the youngest; I_1 – I_3 , sites of incipient organ formation. im, inflorescence meristem; P, leaf primordia in **a** and flower primordia in **k** and **l**. White arrows indicate the vasculature, yellow arrowheads indicate polar PIN1 signal. Stars denote the meristem centre. Scale bars, 50 μ m (**a**, **b**, **i**–**m**); 10 μ m (**c**–**h**).

The *monopteros* (*mp*) and *pinoid* (*pid*) mutants display a defect in organ formation similar to that in *pin1*, that is, they fail to induce flowers but are normal with respect to stem growth and meristem maintenance^{28,29}. Hence, *PIN1*, *MP* and *PID* may function in the same pathway in flower formation. *PID* encodes a protein kinase that has been implicated both in negative regulation of auxin response and in positive regulation of polar auxin transport^{30,31}. *MP* encodes an ARF-type auxin response transcription factor¹¹. While these mutants are similar with respect to their inflorescence phenotype, they can be distinguished by their response to auxin microapplication. *mp* mutant meristems were unresponsive to

1 mM IAA (Fig. 2c), or to 10 mM IAA in the lanolin paste ($n = 25$, data not shown), suggesting a function of MP in the response to auxin. However, the stems of *mp* plants still responded with anisotropic growth to lateral IAA application (Fig. 2d), showing that *mp* plants are not generally auxin-insensitive. *pid* mutant apices, like those of *pin1* mutants, responded to exogenous IAA with the formation of flower primordia (Fig. 2e). Hence, the defect in *pid* is likely to be upstream of auxin action. Apices of *pin1;pid* double mutants resembled *pin1* and responded to IAA with organ formation (Fig. 2f), although at lower frequency than the *pin1* and *pid* single mutants.

Expression of PIN1 in *mp*, *pid* and *pin1*

The mutant phenotypes of *pin1*, *pid* and *mp* show that auxin transport and perception are critical for flower initiation. To see whether the patterning of the inflorescence meristem was also affected in these mutants we analysed PIN1 expression, which exhibits a phyllotactic pattern in wild-type plants³² (Fig. 1b, 1i).

In *mp* mutants, PIN1 was expressed in the L₁ layer, and was localized in a polar fashion (Fig. 2g, h). However, PIN1 was uniformly expressed at the flank of the meristem and lacked a phyllotactic pattern. In the *pid* mutant, PIN1 protein did not accumulate to normal levels, and did not show a phyllotactic pattern (Fig. 2i). PIN1 protein could not be detected in *pin1* mutants;

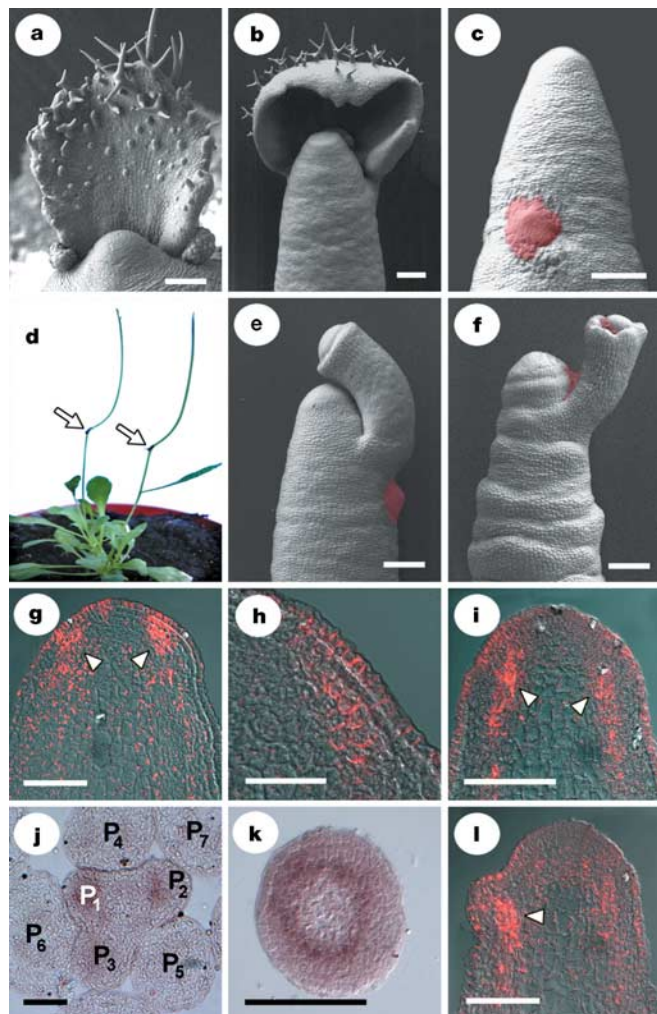


Figure 2 Auxin-induced leaf and flower formation and PIN1 expression in mutants of *Arabidopsis*. **a**, Leaf formation triggered on a vegetative *pin1* meristem by exogenous IAA (66% induction, $n = 24$). **b**, Leaf formation on a *pin1;lfy* double mutant treated with exogenous IAA (20% leaves, $n = 64$ versus 0% leaves in *pin1* single mutant). **c**, Application of IAA to *mp* meristems did not induce organ formation ($n = 45$). **d**, Auxin response (arrows) in the stem of *mp* plants (80% bending, $n = 25$). **e**, Organ formation on *pid* induced with IAA (92% induction, $n = 24$). **f**, Organ formation on *pin;pid* induced with IAA (70% induction, $n = 50$). **g**, Longitudinal section through a *mp* mutant apex, showing PIN1 signal on both sides of the meristem flank (arrowheads), and in the L₁ layer. **h**, Section as in **g** showing polar localization of PIN1 in the L₁ layer. **i**, Longitudinal section through a *pid* mutant apex, showing PIN1 signal at the flank, but weaker and in cells more basal than in *mp* (compare with **g**). **j**, Cross-section of a wild-type inflorescence apex showing PIN1 mRNA accumulation following the phyllotactic pattern. **k**, Cross-section of a *pin1* apex with PIN1 mRNA accumulation in the entire periphery. **l**, Longitudinal section of a *pid* apex with PIN1 induction in a flower primordium triggered by exogenous IAA. White arrowheads indicate PIN1 signal. Lanolin paste in SEM images was pseudo-coloured red in **c**, **e** and **f** for clarity. Scale bars, 100 μ m (**a–f**); 50 μ m (**g, i–l**); 25 μ m (**h**).

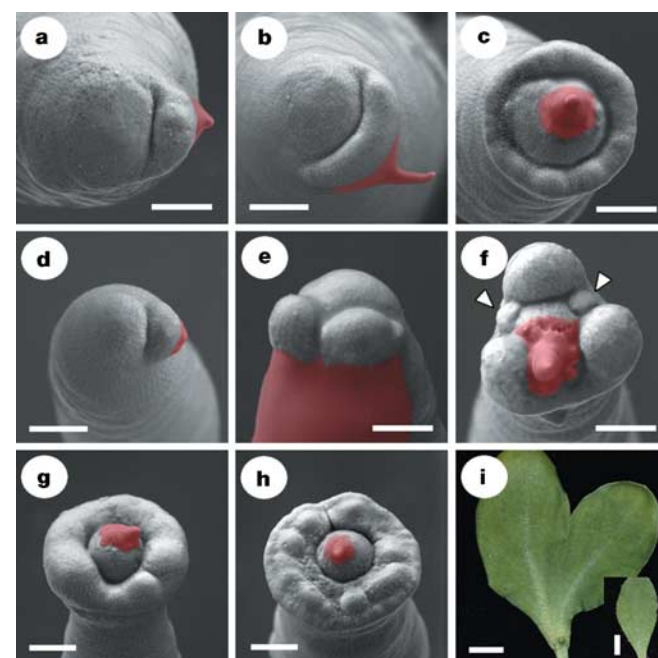


Figure 3 The role of PIN1 in organ separation and delimitation. **a**, Small flower primordium induced by small droplet of exogenous IAA at the flank of a *pin1* meristem (90% induction; $n = 20$). **b**, Wide flower primordium induced by larger droplet of exogenous IAA at the flank of a *pin1* meristem (93% induction; $n = 27$). **c**, Ring-shaped flower primordium induced by IAA application to the meristem centre of a *pin1* meristem (60% complete ring, 40% incomplete ring; $n = 20$). **d**, Small flower primordium induced by a small droplet of exogenous IAA at the flank of a *pid* meristem (88% induction; $n = 25$). **e**, Pair of separated small flower primordia induced by large amount of exogenous IAA at the flank of a *pid* meristem (63% pairs, 21% single organ; $n = 24$). **f**, Multiple small flower primordia induced by IAA application on the meristem centre of a *pid* meristem. Note the formation of a second whorl of organs (arrowheads) in the gaps between primordia of the previous whorl (95% induction; $n = 20$). **g**, Treatment of *pid* as in **f** but with 2,4-D instead of IAA resulted in the formation of ring-shaped flower primordia (50% ring; 6% incomplete ring; $n = 30$). **h**, Treatment of a *pin1;pid* double mutant on the meristem centre resulted in the formation of a ring (50% ring; 20% incomplete ring; $n = 20$). **i**, Fused leaf of a *pin1* mutant, compared to a leaf of *pid* (inset). Scale bars, 100 μ m (**a–h**); 1 mm (**i**).

however, the messenger RNA still accumulated, allowing us to determine the expression pattern of *PIN1* in *pin1* mutants by *in situ* hybridization. In wild-type apices, *PIN1* mRNA was expressed in a phyllotactic pattern similar to the *PIN1* protein (Fig. 2j), but, in *pin1* mutants, *PIN1* mRNA was expressed throughout the periphery of the meristem (Fig. 2k). This shows that *PIN1* expression not only responds to phyllotactic information, but that functional *PIN1* protein is necessary to create such patterns. Treatment of *pid* mutants with IAA resulted in the induction of *PIN1* in the emerging organs (Fig. 2l), as in wild-type organ formation (Fig. 1k). Thus, *PIN1* expression responds to auxin, directly or indirectly.

The uniform expression patterns of *PIN1* at the periphery in *pin1*, *mp* and *pid* are reminiscent of the ring-shaped expression domains of organ and boundary marker genes in the periphery of *pin1* meristems³². Similarly, in the *pid* mutant, the organ marker *LFY* as well as the boundary marker *CUC2* were expressed around the entire periphery (data not shown). Taken together, these results show that *PIN1*, and auxin, play a central role in phyllotactic patterning. On the other hand, *PIN1* responds to phyllotactic patterning information, indicating that phyllotaxis involves a feedback mechanism.

Role of *PIN1* in organ separation and delimitation

During embryogenesis, auxin efflux carrier activity is critical to establish bilateral symmetry and cotyledon separation^{9,10}. To test the role of *PIN1* in organ separation and delimitation in the shoot meristem, we applied lanolin droplets of different sizes containing IAA to the flank of the meristem of *pin1* and *pid* mutants. In *pin1* mutants, the amount of applied IAA determined the size of the induced organs (Fig. 3a, b). When IAA was applied to the tip of the meristem, a ring-shaped organ was induced, indicating that, in the absence of *PIN1* function, IAA could diffuse through the meristem and that the entire periphery of the *pin1* meristem was capable of organ formation (Fig. 3c). In contrast, *pid* mutant apices formed organs of a defined size, and application of larger amounts of IAA increased the number of organs instead of their size (Fig. 3d, e). Application of auxin to the tip induced a whorl of separated organs instead of a ring (Fig. 3f). Interestingly, such whorls were often symmetrically organized, and sometimes, a second whorl of organs was initiated, with the individual organs always being formed between the organs of the previous whorl (Fig. 3f, arrowheads).

It could be argued that auxin placed on the tip of *pid* mutants merely revealed a prepattern present before the treatment. To test this possibility, we treated *pid* apices with 2,4-D, a synthetic auxin analogue which is a poor substrate for the efflux carrier³³ but which can induce organ formation. 2,4-D treatments on the top of the

meristem of *pid* mutants induced ring-shaped primordia (Fig. 3g), similar to those formed on IAA-treated *pin1* mutants (Fig. 3c). Thus, no prepattern had been present at the time of the treatment, which is in agreement with the lack of a visible phyllotactic expression pattern of *PIN1* in *pid* meristems (Fig. 2i). It follows that patterns like those seen in Fig. 3e and f must have been generated after auxin application. We hypothesize that *PIN1*-dependent patterning was triggered by the exogenous auxin and resulted in the generation of separate auxin peaks from an initially even auxin distribution around the periphery of the meristem. According to this scenario, auxin plays an instructive role in the patterning of the apex, and is not just a permissive factor in primordium outgrowth. IAA treatment of *pin1;pid* double mutant meristems on the meristem centre induced the formation of rings similar to those in *pin1* single mutants (Fig. 3h). Thus, *pin1* is epistatic to *pid*, indicating that the patterning induced by IAA in *pid* mutants requires the function of *PIN1*. The defect of *pin1* mutants in flower primordium separation correlates well with defects in delimitation and separation of leaves and cotyledons^{17,23,27,28} (Fig. 3i).

Role of *AUX1* in phyllotaxis

Our data demonstrate a prominent role of the putative efflux carrier *PIN1* in the phyllotaxis of *Arabidopsis*. However, in tomato, pharmacological studies had revealed an additional role for auxin influx carriers in the lateral delimitation of leaves³⁴. Mutations in the putative auxin influx carrier *AUX1* of *Arabidopsis* did not visibly affect leaf formation (data not shown), suggesting that either influx carrier function is not limiting in *Arabidopsis* meristems, or that redundancy of function with the three related *Like-AUX1* genes³⁵ (*LAX* genes) hides the loss-of-function phenotype of *aux1* mutants. To explore these possibilities, we crossed *aux1* into the *pin1* background, reasoning that in the absence of the efflux carrier, the role of *AUX1* could become limiting. When the apices of such double mutants were treated with IAA at the flank, organs were formed that were much wider than in *pin1* single mutants (Fig. 4a, compare with Fig. 3a). Thus, auxin influx carrier activity contributes to organ delimitation, but in the presence of active *PIN1*, the contribution of *AUX1* is masked.

To determine the expression pattern and subcellular localization of *AUX1*, we performed immunofluorescent detection of a haemagglutinin-tagged version of *AUX1* (referred to as HA-*AUX1*)³⁶. In vegetative apices, HA-*AUX1* was detected in the abaxial epidermis of all leaf primordia, and in the meristem *L*₁ layer (Fig. 4b), where the signal was stronger at the flank compared to the meristem centre (Fig. 4c). On the basis of this expression pattern, we think the function of *AUX1* could be to promote intracellular auxin

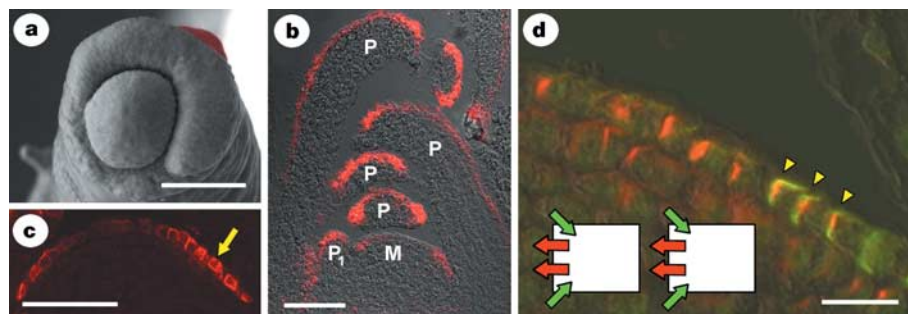


Figure 4 Localization and function of *AUX1* in the meristem. **a**, In the *pin1;aux1* double mutant, small droplets of auxin induced the formation of oversized flower primordia (compare to Fig. 3a) (46% ring; 11% incomplete ring; 17% local organ; *n* = 120). **b**, Expression of HA-tagged *AUX1* in a 6-week-old vegetative meristem. Signal was strongest in the abaxial epidermis of *P*₁ and older leaf primordia (*P*). **c**, At 6 weeks, *AUX1* was expressed in the *L*₁ cells of the meristem with the strongest expression at the

periphery. Localization was to the plasmalemma with hints of subcellular polarity (arrow). **d**, Double labelling on a median section of a 4-week-old plant with antibodies against HA-*AUX1* (green) and *PIN1* (red). Localization of *AUX1* and *PIN1* partially overlapped (arrowheads). Inset, possible net auxin fluxes inferred from the localization and relative levels of *AUX1* and *PIN1*. M, meristem; P, primordia. Scale bars, 50 μm (**a–c**); 10 μm (**d**).

accumulation in the L_1 layer and abaxial epidermis.

In 6-week-old plants, AUX1 signal was clearly localized to the plasmalemma, and showed only hints of subcellular polarity (Fig. 4c). During early vegetative development, subcellular localization of HA-AUX1 signal was more evident, and frequently coincided on the same side of the cells as PIN1 signal, that is, on the distal side of the cells (Fig. 4d). However, while the PIN1 signal was restricted to the anticlinal membrane towards the meristem summit, AUX1 signal was also observed at the edges of the cells and in the outer and inner periclinal membrane of the L_1 cells (Fig. 4d).

The coexpression of AUX1 and PIN1 in the L_1 layer could point to a common function in auxin trafficking in these cells. We propose that AUX1 restricts auxin to the outer cell layer (by counteracting loss of auxin via diffusion to inner layers), and that PIN1 promotes its directional transport towards the tip of the meristem and of leaf primordia. After extrusion of auxin by PIN1, AUX1 could retrieve apoplastic auxin before it diffuses away to the inner layers. A parallel could be drawn with the vertebrate brain, where the neurotransmitter glutamate is removed from the synaptic cleft by re-uptake into presynaptic terminals, in order to return to a low resting concentration after exocytotic release of glutamate³⁷.

Discussion

Phyllotaxis is one of the most conspicuous examples of patterning in plants, and our results reveal a central role of auxin in its molecular regulation. The accumulated evidence clearly indicates an instructive rather than a permissive role for auxin in phyllotaxis. First, genetic and pharmacological interference with auxin transport or response leads to a wide range of phyllotactic patterning defects at all stages of development. Second, the phyllotactic expression patterns of several regulatory genes were abolished in *pin1*, as well as in *pid* and *mp*. Third, if *PIN1* just transmitted phyllotactic information from an upstream regulator, then the *PIN1* mRNA should be expressed in the normal phyllotactic pattern in *pin1*. Thus, the lack of phyllotactic expression of *PIN1* in the *pin1* mutant shows that *PIN1* not only responds to the phyllotactic signal (auxin) but also itself creates phyllotactic pattern. Fourth, exogenous auxin application does not only restore organogenesis, but the observed effects depend on the particular mutant, on the concen-

tration of auxin and on the site of application. Fifth, application of the natural auxin IAA, but not the synthetic auxin 2,4-D (which is a poor substrate for the efflux carrier), can induce the PIN1-dependent generation of *de novo* phyllotactic pattern on the *pid* mutant (Fig. 3). Sixth, the localization of the putative auxin transporters is highly informative and predicts auxin distribution in a dynamic pattern, with auxin accumulating each plastochron at the site of incipient organ initiation.

Based on the experimental data, we propose a model for the regulation of phyllotaxis in *Arabidopsis* (Fig. 5). Auxin is acropetally transported towards the meristem. In the meristem, auxin becomes absorbed by the primordia, which function as sinks. As a result, auxin is depleted from the surroundings of the primordia and reaches the organogenetic peripheral zone only at a certain minimal distance from P_1 and P_2 (corresponding to the I_1 position). Auxin accumulation at I_1 induces a new primordium, which, in the course of the plastochron, will grow out and itself become a sink for auxin. This mechanism represents a combination of positive feedback (auxin accumulation) and lateral inhibition (withdrawal of auxin from adjacent tissues), that could conceptually be compared with the short-range activator and the long-range inhibitor in reaction-diffusion mechanisms⁸.

The described system would require dynamic regulation of PIN1 during the course of the plastochron. Indeed, cellular redistribution of the PIN1 protein can be extremely rapid^{21,38} and could easily allow for the 137° shift in orientation required during each plastochron. In the root, the subcellular localization of PIN1 is dynamically regulated by GNOM, which is a component of the vesicular cycling mechanism^{24,39}. A similar mechanism could be active in the meristem. We also note that genes such as *PINHEAD/ZWILLE* are expressed in a phyllotactic pattern well below the meristem, and might also contribute to phyllotactic patterning⁴⁰. However, the lack of patterned expression of all genes tested in *pin1*, *pid* and *mp* show that phyllotactic patterning in the *Arabidopsis* inflorescence cannot proceed independently of auxin. Essential downstream effectors that are required to translate auxin-based patterning information into actual growth patterns are likely to include factors such as *CUP-SHAPED COTYLEDON1* and 2 (*CUC1* and *CUC2*)^{12,41} for organ separation, and expansin for organ outgrowth⁴².

Our model accounts for the reiterativity and the stability of organ positioning. But how are the divergence angles determined? We speculate that if only P_1 is effective as a sink, a distichous phyllotaxis (divergence angle 180°) would result. Higher-order phyllotactic systems will result from the contribution of increasing numbers of primordia. Important parameters influencing relative sink strength are likely to be the size of the central zone relative to the organogenetic zone (plastochron ratio) and the growth rate of the apex along the apical-basal axis⁴³. In addition, the timing of the transition of primordia from auxin sinks to auxin sources will influence phyllotaxis. Our work addresses the maintenance of phyllotactic pattern. The challenge now is to elucidate how phyllotactic patterns are generated *de novo*, and how they change under natural conditions, for example when dicotyledonous plants undergo the transition from decussate to spiral phyllotaxis. □

Methods

Plant material and growth conditions

Arabidopsis thaliana plants were grown on soil under short-day conditions (8 h light, 60% humidity, 20–22°C day temperature, irradiance 150 $\mu\text{E m}^{-2} \text{s}^{-1}$). HA-AUX1 was generated as described previously³⁶. The *pin1-1* allele (Enkheim ecotype, provided by the ABRC, Ohio State University), was used for all experiments, except for treatments of vegetative apices, where the *pin1-6* allele (Wassilewskaya ecotype) was used³², and for the *pin1;pid* double mutant which was generated with the alleles *pin1-6* and *pid-14* (Wassilewskaya ecotype), respectively. *pid-9* (Columbia ecotype), which was used for all other *pid*-experiments, was a gift from S. Christensen³⁰, and *mp*^{G12} (Columbia ecotype) was provided by T. Berleth¹¹. For the *aux1;pin1* double mutant the allele *aux1-22* (Columbia ecotype) was used²². *Lfy-6* (Columbia ecotype) was a gift from D. Weigel²⁶. All mutant alleles used were, as shown by phenotypical analysis and molecular characterization, very strong or null.

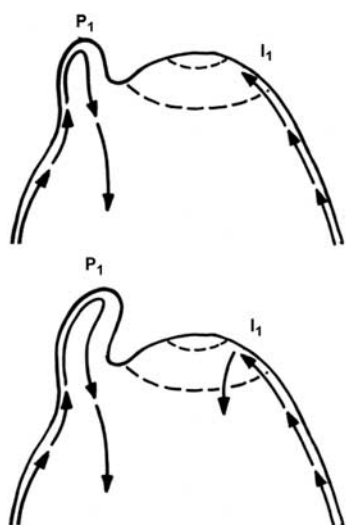


Figure 5 Model for the role of polar auxin transport in phyllotaxis. Schematic representation of an apex in longitudinal section through P_1 and I_1 at an early (top) and a later stage (bottom) of incipient primordium formation. Polar auxin flux is indicated with arrows. Top, acropetal auxin flux is diverted by P_1 preventing auxin accumulation on the left flank of the meristem, while auxin can reach the right flank (I_1). Bottom, accumulation of auxin at I_1 promotes primordium formation, and establishment of a new auxin sink.

In situ detection of mRNA and proteins

In situ hybridizations were carried out with digoxigenin-labelled riboprobes³². For protein detections, *Arabidopsis* apices of 4- to 6-week-old plants were dissected and fixed overnight at -20°C in a methanol-acetic acid solution (3:1). Fixed tissue was embedded in paraplast (Sigma) and sectioned ($8\text{ }\mu\text{m}$). After removal of paraplast and rehydration, immunofluorescent detection of AUX1 and PIN1 was carried out essentially as described^{15,36}, using primary mouse antibody against HA (BAbCO, Richmond, CA, 1:200), primary rabbit anti-PIN1 antibody²³, and CyTM3-conjugated goat secondary antibody (Jackson ImmunoResearch, 1:200). In the case of double labellings, Alexa Fluor 488-conjugated goat secondary antibody (Molecular Probes) was used to detect HA-AUX1. Slides were mounted using Faramount medium (Dako) after several washings with PBS pH 7.4. Fluorescence was detected with a laser-scanning confocal microscope (Zeiss Axiovert 100M; Zeiss LSM510) equipped with a Zeiss Plan-Neofluar $40\times 1.3\text{ NA}$ oil immersion objective. Fluorescence in Fig. 1i was recorded with a Zeiss Axioskop 2 (non-confocal). The fluorescence images were underlaid with the phase-contrast images.

Microapplications and microscopy

Microapplications with 1 mM IAA and 1 mM 2,4-D were carried out as described¹⁷. For scanning electron microscopic analysis, apices were viewed with a S-3500N variable pressure scanning electron microscope from Hitachi, equipped with a cool stage. Lanolin paste in digital images was pseudocoloured for clarity.

Received 25 June; accepted 29 September 2003; doi:10.1038/nature02081.

1. Steeves, T. A. & Sussex, I. M. *Patterns in Plant Development* (Cambridge Univ. Press, Cambridge, UK, 1989).
2. Reinhardt, D. & Kuhlemeier, C. in *Meristematic Tissues in Plant Growth and Development* (eds McManus, M. T. & Veit, B. E.) 172–212 (Sheffield Academic, Sheffield, 2002).
3. Mitchison, G. J. Phyllotaxis and the Fibonacci series. *Science* **196**, 270–275 (1977).
4. Jean, R. V. *Phyllotaxis: A Systematic Study in Plant Morphogenesis* (Cambridge Univ. Press, Cambridge, UK, 1994).
5. Snow, M. & Snow, R. A theory of regulation of phyllotaxis based on *Lupinus albus*. *Phil. Trans. R. Soc. Lond. Ser. B* **244**, 483–513 (1962).
6. Green, P. B. Expression of form and pattern in plants - A role for biophysical fields. *Semin. Cell Dev. Biol.* **7**, 903–911 (1996).
7. Schoute, J. C. Beiträge zur Blattstellungslehre. *Résumé des Travaux Botaniques Néerlandais* **10**, 153–325 (1913).
8. Meinhardt, H. in *Positional Controls in Plant Development* (eds Barlow, P. W. & Carr, D. J.) 1–32 (Cambridge University Press, Cambridge, UK, 1984).
9. Liu, C.-M., Xu, Z.-H. & Chua, N.-H. Auxin polar transport is essential for the establishment of bilateral symmetry during early plant embryogenesis. *Plant Cell* **5**, 621–630 (1993).
10. Hadfi, K., Speth, V. & Neuhaus, G. Auxin-induced developmental patterns in *Brassica juncea* embryos. *Development* **125**, 879–887 (1998).
11. Hardtke, C. S. & Berleth, T. The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor mediating embryo axis formation and vascular development. *EMBO J.* **17**, 1405–1411 (1998).
12. Aida, M., Vernoux, T., Furutani, M., Traas, J. & Tasaka, M. Roles of *PIN-FORMED1* and *MONOPTEROS* in pattern formation of the apical region of the *Arabidopsis* embryo. *Development* **129**, 3965–3974 (2002).
13. Jürgens, G. Apical-basal pattern formation in *Arabidopsis* embryogenesis. *EMBO J.* **20**, 3609–3616 (2001).
14. Sabatini, S. *et al.* An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99**, 463–472 (1999).
15. Friml, J. *et al.* AtPIN4 mediates sink-driven auxin gradients and root patterning in *Arabidopsis*. *Cell* **108**, 661–673 (2002).
16. Mattsson, J., Ckurshumova, W. & Berleth, T. Auxin signaling in *Arabidopsis* leaf vascular development. *Plant Physiol.* **131**, 1327–1339 (2003).
17. Reinhardt, D., Mandel, T. & Kuhlemeier, C. Auxin regulates the initiation and radial position of plant lateral organs. *Plant Cell* **12**, 507–518 (2000).
18. Kuhlemeier, C. & Reinhardt, D. Auxin and phyllotaxis. *Trends Plant Sci.* **6**, 187–189 (2001).
19. Lomax, T. L., Muday, G. K. & Rubery, P. H. in *Plant Hormones: Physiology, Biochemistry and Molecular Biology* (ed. Davies, P. J.) 509–530 (Kluwer Academic, Dordrecht, 1995).

20. Friml, J. & Palme, K. Polar auxin transport—old questions and new concepts? *Plant Mol. Biol.* **49**, 273–284 (2002).
21. Friml, J., Wisniewska, J., Benková, E., Mendgen, K. & Palme, K. Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* **415**, 806–809 (2002).
22. Bennett, M. J. *et al.* *Arabidopsis* AUX1 gene: A permease-like regulator of root gravitropism. *Science* **273**, 948–950 (1996).
23. Gälweiler, L. *et al.* Regulation of polar auxin transport by AtPIN1 in *Arabidopsis* vascular tissue. *Science* **282**, 2226–2230 (1998).
24. Steinmann, T. *et al.* Coordinated polar localization of auxin efflux carrier PIN1 by GNOM ARF GEF. *Science* **286**, 316–318 (1999).
25. Sachs, T. The control of patterned differentiation of vascular tissues. *Adv. Bot. Res.* **9**, 151–262 (1981).
26. Weigel, D., Alvarez, J., Smyth, D. R., Yanofsky, M. F. & Meyerowitz, E. M. *LEAFY* controls floral meristem identity in *Arabidopsis*. *Cell* **69**, 843–859 (1992).
27. Okada, K., Ueda, J., Komaki, M. K., Bell, C. J. & Shimura, Y. Requirement of the auxin polar transport-system in early stages of *Arabidopsis* floral bud formation. *Plant Cell* **3**, 677–684 (1991).
28. Bennett, S. R. M., Alvarez, J., Bossinger, G. & Smyth, D. R. Morphogenesis in *pinoid* mutants of *Arabidopsis thaliana*. *Plant J.* **8**, 505–520 (1995).
29. Przemeck, G. K. H., Mattsson, J., Hardtke, C. S., Sung, Z. R. & Berleth, T. Studies on the role of the *Arabidopsis* gene *MONOPTEROS* in vascular development and plant cell axialization. *Planta* **200**, 229–237 (1996).
30. Christensen, S. K., Dagenais, N., Chory, J. & Weigel, D. Regulation of auxin response by the protein kinase PINOID. *Cell* **100**, 469–478 (2000).
31. Benjamins, R., Quint, A., Weijers, D., Hooykaas, P. & Offringa, R. The PINOID protein kinase regulates organ development in *Arabidopsis* by enhancing polar auxin transport. *Development* **128**, 4057–4067 (2001).
32. Vernoux, T., Kronenberger, J., Grandjean, O., Laufs, P. & Traas, J. *PIN-FORMED1* regulates cell fate at the periphery of the shoot apical meristem. *Development* **127**, 5157–5165 (2000).
33. Delbarre, A., Muller, P., Imhoff, V. & Guern, J. Comparison of mechanisms controlling uptake and accumulation of 2,4-dichlorophenoxy acetic acid, naphthalene-1-acetic acid, and indole-3-acetic acid in suspension-cultured tobacco cells. *Planta* **198**, 532–541 (1996).
34. Stieger, P. A., Reinhardt, D. & Kuhlemeier, C. The auxin influx carrier is essential for correct leaf positioning. *Plant J.* **32**, 509–517 (2002).
35. Parry, G. *et al.* Quick on the uptake: Characterization of a family of plant auxin influx carriers. *J. Plant Growth Regul.* **20**, 217–225 (2001).
36. Swarup, R. *et al.* Localization of the auxin permease AUX1 suggests two functionally distinct hormone transport pathways operate in the *Arabidopsis* root apex. *Genes Dev.* **15**, 2648–2653 (2001).
37. Danbolt, N. C. Glutamate uptake. *Prog. Neurobiol.* **65**, 1–105 (2001).
38. Geldner, N., Friml, J., Stierhof, Y. D., Jürgens, G. & Palme, K. Auxin transport inhibitors block PIN1 cycling and vesicle trafficking. *Nature* **413**, 425–428 (2001).
39. Geldner, N. *et al.* The *Arabidopsis* GNOM ARF-GEF mediates endosomal recycling, auxin transport, and auxin-dependent plant growth. *Cell* **112**, 219–230 (2003).
40. Lynn, K. *et al.* The *PINHEAD/ZWILLE* gene acts pleiotropically in *Arabidopsis* development and has overlapping functions with the *ARGONAUTE1* gene. *Development* **126**, 469–481 (1999).
41. Aida, M., Ishida, T., Fukaki, H., Fujisawa, H. & Tasaka, M. Genes involved in organ separation in *Arabidopsis*: An analysis of the *cup-shaped cotyledon* mutant. *Plant Cell* **9**, 841–857 (1997).
42. Fleming, A. J., McQueen-Mason, S., Mandel, T. & Kuhlemeier, C. Induction of leaf primordia by the cell wall protein expansin. *Science* **276**, 1415–1418 (1997).
43. Schwabe, W. W. in *Positional Controls in Plant Development* (eds Barlow, P. W. & Carr, D. J.) 403–440 (Cambridge Univ. Press, Cambridge, UK, 1984).

Acknowledgements We thank J. Moore, J. Stuurman and S. Zeeman for critical reading of the manuscript, and T. Vernoux for generating *pin1;pid* double mutants and H. Morin for supporting experiments. This work was supported by grants from the Swiss National Science Foundation and from the European Union.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.K. (cris.kuhlemeier@ips.unibe.ch).

A collimated, high-speed outflow from the dying star V Hydrae

R. Sahai¹, M. Morris², G. R. Knapp³, K. Young⁴ & C. Barnbaum⁵

¹Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA

²Division of Astronomy, Department of Physics and Astrophysics, UCLA, Los Angeles, California 90095, USA

³Princeton University, Department of Astrophysical Sciences, Princeton, New Jersey 08544, USA

⁴Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

⁵Valdosta State University, Department of Physics, Astronomy and Geosciences, Valdosta, Georgia 31698, USA

Stars with masses in the range 1–8 solar masses ($M_{\odot}$) live ordinary lives for $\sim 10^9$ – 10^{10} years, but die extraordinary deaths. First, during their death throes as asymptotic giant branch (AGB) stars they eject, over 10^4 – 10^5 years, half or more of their mass in slowly expanding, spherical winds, and then, in a short (a few

100–1,000 years) and poorly understood phase, they are transformed into aspherical planetary nebula. Recent studies support the idea that high-speed, jet-like flows play a crucial role in this transformation¹. Evidence for such outflows is indirect, however; this phase is so short that few nearby stars are likely to be caught in the act. Here we report the discovery of a newly launched, high-speed jet-like outflow in the nearby AGB star, V Hydrae. We have detected both proper motions and ongoing evolution in the jet. These results support a model in which the jet is driven by an accretion disk around an unseen, compact companion. We also find a central, dense equatorial disk-like structure which may enable and/or enhance the formation of the accretion disk.

Evidence for the presence of mild bipolarity and fast-moving gas in the circumstellar envelope of V Hya (a well-studied, optically-bright, cool carbon star, that is, the C/O abundance ratio in its atmosphere is greater than unity) first came from molecular-line (CO $J = 1-0$) millimetre-wave observations². Spectroscopy of the CO vibration-rotation lines at $4.6 \mu\text{m}$ then showed the presence of a very-high-velocity, kinematically complex outflow³ having at least six velocity components from $V_{\text{exp}} = 70$ to 120 km s^{-1} (ref. 4). CO $J = 1-0$ and $2-1$ maps allowed modelling of the circumstellar environment in V Hya as a fast ($>50 \text{ km s}^{-1}$) bipolar outflow and

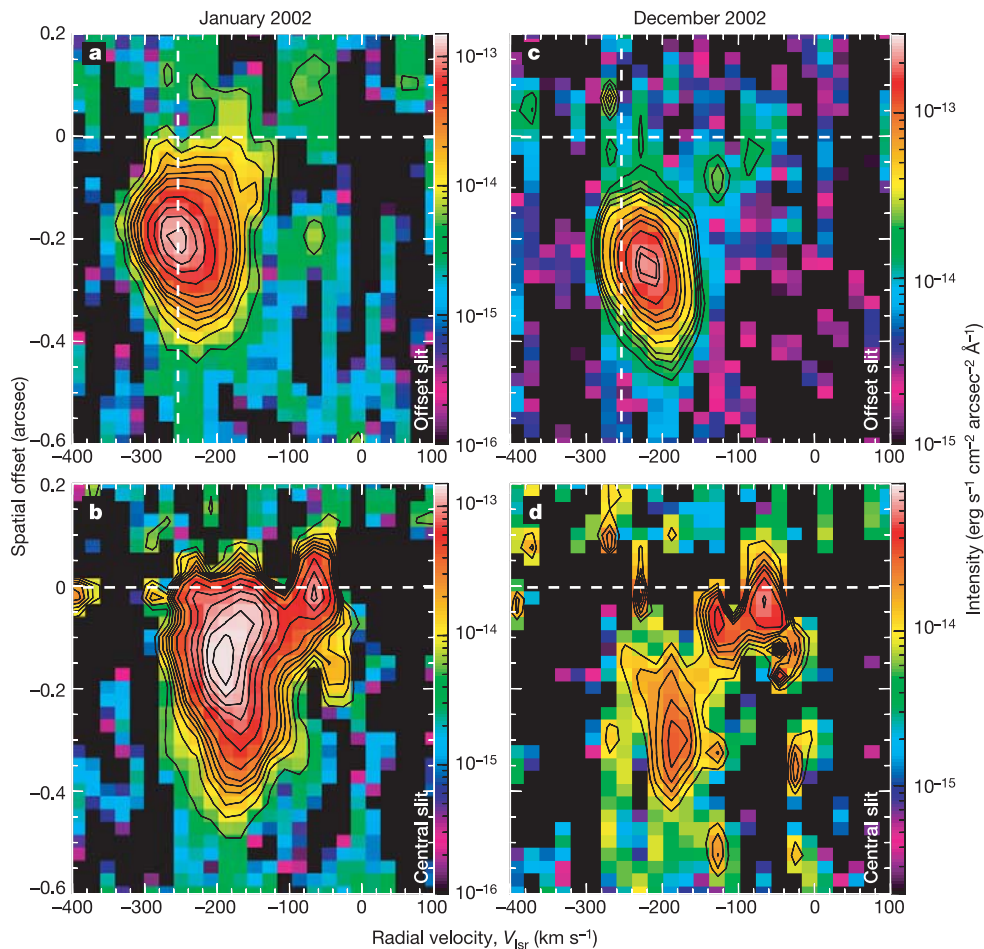


Figure 1 Optical line emission from a high-velocity, jet-like, knotty outflow in the carbon star V Hya. The long-slit spectra show [S II] $\lambda = 4,069.7$ line emission, and were obtained with STIS using the G430M grating, during January 2002 (**a**, **b**) and December 2002 (**c**, **d**). Three (five) adjacent slits $0.2''$ ($0.1''$) wide spaced apart by $0.2''$ ($0.1''$), aligned at a position angle $= -89.9^\circ$ (-92.9°) were used for the January 2002 (December 2002) observations (east is towards negative offsets in the panels). Spectra observed through the central (**b**, **d**) and the first adjacent slit offset to the south (**a**, **c**) from the star, are shown. The location of the star along the slit (marked by the horizontal dashed line) is determined

from the continuum background, which has been subtracted from the spectra shown here. The [S II] emission is blue-shifted (vertical dashed line shows $V_{\text{lsr}} = -255 \text{ km s}^{-1}$, the systemic radial velocity is $V_{\text{lsr}} = -17.5 \text{ km s}^{-1}$) and peaks east of the star (negative offsets). In January 2002, the high-velocity emission was detected only in the $-0.2''$ (that is, south) offset slit (**a**) and the central slit (**b**); whereas in December 2002, it was strongly present only in the $-0.1''$ offset slit (**c**), significantly weaker in the central slit (**d**), and extremely faint or undetectable in the remaining slits (standard pipeline-processed data have been used; a smaller number of 'bad' pixels have been corrected by interpolation.)

a dense, slowly expanding, equatorial toroid⁵. High signal-to-noise CO $J = 2-1$ and $3-2$ line profiles showed wide wings, indicating the presence of a very fast ($\geq 200 \text{ km s}^{-1}$) molecular outflow⁶. V Hya was also found to have highly blue-shifted (-155 km s^{-1} relative to the star) emission lines in the blue-violet spectrum, interpreted as arising in shocks generated by a high-velocity flow interacting with dense, slowly expanding circumstellar material⁷.

The fast outflows seen in V Hya are not typical of AGB stars. But analysis of light curve data confirm that V Hya's primary period variability of 529 days (it also displays a secondary period of 6,160 days) is consistent with its classification as a Mira variable⁸. Furthermore, V Hya's high luminosity ($\sim 10^4 L_{\odot}$)⁸ is typical of AGB stars, and it is located in the region of the Infrared Astronomical Satellite (IRAS) colour-colour diagram populated by AGB stars⁶. Thus, V Hya's AGB classification is fairly secure. However, the outer layers of its atmosphere are rotating anomalously rapidly (revealed by extensive monitoring of the photospheric velocity broadening from high-resolution optical spectra) suggesting the presence of a close companion which may have spun up V Hya's outer envelope⁹.

Our long-slit spectroscopy of V Hya with the Space Telescope Imaging Spectrograph (STIS) on board the Hubble Space Telescope (HST) now provides a first glimpse into the spatio-kinematic structure of the high-velocity outflow and its ongoing evolution. We find an emission-line blob, particularly prominent in the forbidden lines characteristic of fast shocks in molecular regions (such as Herbig-Haro objects), very close to but offset from the star (Fig. 1). The blob has a high negative radial velocity, about -240 km s^{-1} relative to the systemic velocity -17.5 km s^{-1} (V_{lsr}), and second-epoch observations show that it has moved further away from the star (the apparent proper motion, μ , is about $0.065'' \text{ yr}^{-1}$, along the east-west direction). Thus we are probably seeing a directed, very fast outflow.

Simple arguments related to the relative fluxes of the [S II]

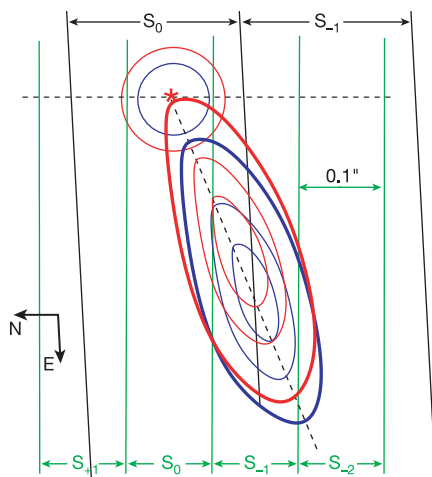


Figure 2 A geometrical model of the high-velocity knotty jet and the central continuum source in V Hya. The figure shows model intensity contours for the geometrical projection onto the plane of the sky of the high-velocity emission feature and the central continuum source for epoch 1 (in red) and epoch 2 (in blue). The centre of the continuum source is marked with an asterisk. All elements of the diagram are roughly correct in relative scale. For epoch 1 (2), the black (green) vertical lines represent the edges of the central (S_0) and adjacent slits S_{-1} (S_{+1} , S_{-1} and S_{-2}); there is a 3° misalignment of the slits between the two epochs. For epoch 1 (2), the size (full-width at half-maximum, FWHM, for assumed gaussian shape) of the continuum source, 120 (81) milliarcsec, and its offset from the centre of the central slit, 20 (4) milliarcsec, are derived from the observed continuum fluxes measured in slits, S_0 , S_{+1} and S_{-1} . The position and opening angles of the blob (assumed to be axially symmetric) are constrained by the relative [S II] emission-line fluxes observed in the different slits (see Fig. 1) at the two epochs, whereas its radial extent is constrained by its observed size along the different slits. The blob is assumed to move along a radial direction from epoch 1 to 2 in order to produce the observed proper motion.

emission line and the continuum in the different slits at both epochs, and the observed proper motion of the blob, allow us to roughly constrain the size and opening angle of the high-velocity outflow and its location relative to the central continuum source. Such a model is shown in Fig. 2. Although careful analysis is necessary for quantitative conclusions, there is unmistakable and significant reduction in the blob radial velocity from January–December 2002 as seen in the offset slits which sample the central, brightest region of the blob at both epochs (Fig. 1). This reduction in the blob velocity most probably results from the deceleration of an underlying fast jet, which produces the observed high-velocity blob, by its interaction with the ambient circumstellar medium. By modelling this deceleration using detailed hydrodynamic simulations of jets interacting with AGB envelopes¹⁰, we can begin to understand the physical properties of this jet.

If we assume that the decrease in the blob velocity is linear, we can use the observed proper motion and the assumed distance, $D = 400 \text{ pc}$ (refs 5, 6, 8), to estimate the inclination angle, i , of the fast outflow to the line of sight, from $\tan i = 2D\mu/(V_{r,1} + V_{r,2})$, where $V_{r,1}$ (or $V_{r,2}$) is the radial velocity of the blob relative to the star at epoch 1 (or 2). We find $i = 20^\circ$, which is comparable to the value of 30° derived for the bipolar CO outflow from millimetre-wave data⁵. The flow is quite collimated, with an opening angle of $\sim (20^\circ-30^\circ) \sin i$, that is, about $7^\circ-10^\circ$, and is directed along position angle $\approx 110^\circ$.

The observed proper motion gives an expansion timescale, t_{exp} , at the peak of the emission in January 2002 of $< 3 \text{ yr}$ (when it was offset by $0.2''$ from the centre), independent of the source distance and outflow inclination. This timescale is small compared to the roughly 12-yr period over which the high-velocity [S II] emission has been

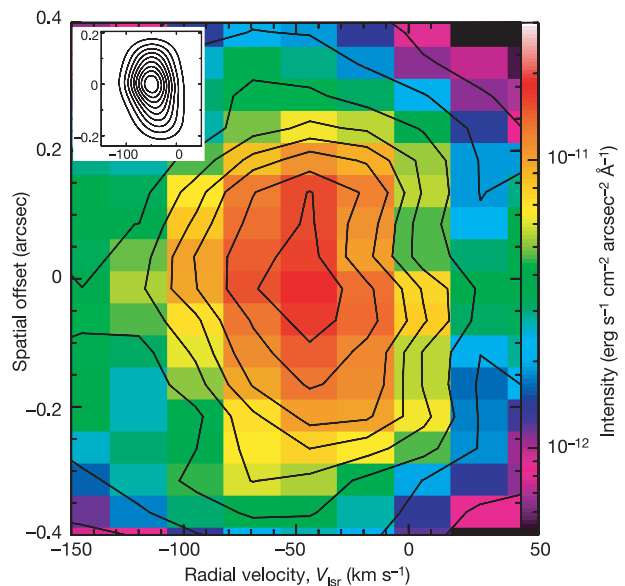


Figure 3 Observations and geometrical model of a central, dense, equatorial disk in V Hya. The figure shows the STIS H α spectrum of V Hya, obtained through a $0.2''$ wide slit, oriented east-west, and offset from the centre $0.2''$ to the north. The H α position-velocity profile is most simply interpreted as resulting from the scattering of H α photons (emitted by a compact, central outflow source) from the surface of an expanding disk. Support for this interpretation is provided by a quantitative (but preliminary) spatio-kinematic model (shown in inset): (1) the disk is co-axial with the high-velocity outflow and is inclined to the line-of-sight by 60° with its east-half (negative offsets) tilted away from us; has a diameter of $0.5''$, is flared such that its periphery is $0.045''$ above its midplane, and its expansion velocity, $V_{\text{exp}}(r) = 10 \text{ km s}^{-1} (r/0.1'')$; and (2) the H α emitting source is assumed to have a radial velocity (V_{lsr}) of -50 km s^{-1} . Model has been convolved with the instrumental spatial ($0.07''$) and spectral resolution (about 70 km s^{-1}). The axis labels in the inset are identical to the ones used for the main figure. The physical parameters of the disk derived from this preliminary model are only rough estimates.

observed in V Hya from the ground (first detected in December 1989, and then monitored until 2000 with no gaps larger than a year^{7,11}). The lack of 'older' emitting material in our data suggests that the [S II] intensity in the outflow decreases rapidly as it expands outwards, indicating that the [S II] emission has been episodically replenished, presumably by a succession of ejections of the sort we have observed.

The blob is seen in a few specific forbidden emission lines in the blue-violet region (where the intrinsic stellar continuum is relatively weak). The strongest of these is the [S II] wavelength $\lambda = 4,069.7$, $\lambda = 4,077.5$ doublet; weak extended emission is also seen, for example, in the H δ and [Fe II] $\lambda = 4,245.2$, $\lambda = 4,288.6$ lines. Other typical line emissions from shocked gas (such as H α , [N II] $\lambda = 6,585.3$, and [S II] $\lambda = 6,718.3$, $\lambda = 6,732.7$) are not seen because the central continuum source is very bright in the red, producing extended surface brightness due to scattered light from circumstellar dust as well as the instrumental point spread function, which overwhelms the line emission. H α , the strongest line pre-

dicted by planar-shock models for shock speeds in the range 220–260 km s⁻¹ and covered by our observations, has an intensity comparable to [S II] $\lambda = 4,072$ (ref. 12), but the scattered continuum at wavelengths in the vicinity of H α is 25–40 times stronger.

Given that the millimetre-wave CO observations of V Hya show the presence of a bipolar outflow (oriented roughly east–west), and that several other bipolar objects (for example, IRAS 16342–3814 and Hen 3–1475) show a point-symmetric morphology, we assume that the optically observed (collimated, high-velocity) outflow is bidirectional. Hence the absence of the expected red-shifted emission from the receding component in the [S II] spectra suggests the presence of an obscuring, dusty structure. This structure is likely to be a tilted equatorial disk (that is, the inner part of the dense, equatorial toroid described in ref. 5), as suggested by our STIS H α profiles (obtained only in the January 2002 epoch). The signature of the disk is best seen in the spectrum taken through an offset slit (Fig. 3) where the contribution of the central continuum source is minimized: spatially extended H α emission, red-shifted relative to the emission at the centre, can be seen at both positive and negative spatial offsets from the stellar continuum. A simple interpretation of this profile is that it results from H α photons emitted from a compact outflow from the central source (because the emission at the centre is blue-shifted relative to the systemic velocity) and scattered off the disk towards us (supported by a quantitative geometric model of a slowly expanding inclined, flared, scattering disk). Direct support for an expanding, tilted, dense disk-like structure comes from an interferometric CO $J = 1-0$ map of V Hya, obtained using the Owen's Valley Radio Observatory in April 1998 (Fig. 4).

What is the origin of the jet and disk in V Hya? The launching and collimation of astrophysical jets is not fully understood; however there is general agreement that most jets are driven from the centres of accretion disks¹³. In V Hya, a detached binary companion may capture material from the red giant wind into an accretion disk which may then power a collimated flow^{14,15}. Knapp *et al.*⁸ suggest that the presence of such a detached binary companion (separation 3.5×10^{14} cm or $\sim 10R_*$, where R_* is the radius of the primary) with an attached thick cloud of obscuring dust can explain the eclipse-like character of V Hya's 6,000-day light curve. Although the expansive kinematics and large sizes of the disk-like structures seen in H α and CO emission imply that these are not the result of accretion, they may be due to a recent phase of equatorially enhanced mass loss in V Hya, which could enhance the accretion process. Alternatively, V Hya may be a contact binary in a common-envelope configuration⁹ which may produce (through tidal interaction) a rapidly rotating, magnetized stellar atmosphere; the latter could power a jet and/or generate a dense equatorial toroid¹⁶.

The jets found in V Hya and a small handful of other AGB (or very early post-AGB) stars^{17–19} show speeds of about a few 100 km s⁻¹. This indicates that the jets are not launched by the AGB stars themselves (whose escape speeds are low), but by compact (main-sequence or white dwarf) companions. In addition to evolved stars, jets are found in a wide variety of object classes, including symbiotic stars²⁰ and young stellar objects²¹. In particular, the jet in the young stellar object HH34 shows speed and deceleration²² similar to those found for the jet in V Hya. Recently, Soker²³ has found interesting similarities between X-ray-deficient bubbles in galaxy clusters and planetary nebulae, and postulated that the formation mechanism for the bubbles in these two object classes is the same, namely, collimated fast winds. Thus, our detailed study of the V Hya jet should help in improving our general understanding of astrophysical jets and their operation. □

Received 2 July; accepted 29 September 2003; doi:10.1038/nature02086.

1. Sahai, R. & Trauger, J. T. Multipolar bubbles and jets in low-excitation planetary nebulae: toward a new understanding of the formation and shaping of planetary nebulae. *Astron. J.* **116**, 1357–1366 (1998).
2. Kahane, C., Maizels (Barnbaum), C. & Jura, M. The bipolar outflow from the rotating carbon star, V Hydrae. *Astrophys. J.* **328**, L25–L28 (1988).
3. Sahai, R. & Wannier, P. Discovery of very high velocity outflow in V Hydra—Wind from an accretion

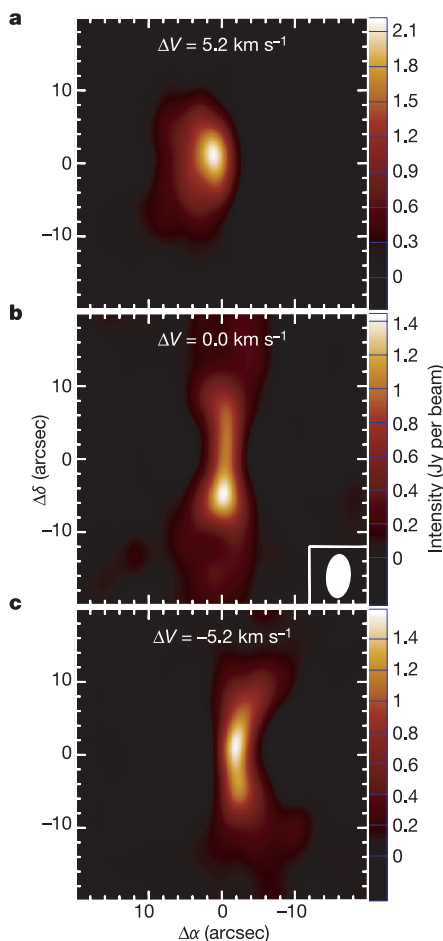


Figure 4 Maps of the CO $J = 1-0$ emission from V Hya. Each panel shows data integrated over a small radial velocity interval 5.2 km s⁻¹ wide and centred on the indicated radial velocity offset from the systemic velocity. Panel **b** is effectively a cross-sectional cut of the sky-plane through the circumstellar envelope. These maps are consistent with the presence of a dense disk-like structure, oriented north–south with its east-half (which appears red-shifted, **a**) tilted away from us and west-half (which appears blue-shifted, **c**) tilted towards us (consistent with the orientation/tilt of the H α disk). The data were obtained with the Owens Valley Radio Observatory Millimeter Array (OVRO), using six antennas and four tracks, two each in the E (equatorial) and the L (low resolution) configurations, giving a synthesized beam of $6.08'' \times 3.41''$ (inset to **b**). Single sideband system temperatures were ~ 500 – 700 K, depending on source elevation, and the precipitable water vapour was typically about 6 mm. Passband calibration was done with 3C 273 and the instrument's noise source. Phase and flux calibration was performed using QSO1045–188.

- disk in a binary? *Astron. Astrophys.* **201**, L9–L12 (1988).
4. Sugerman, B., Sahai, R. & Hinkle, K. in *CO: Twenty-Five Years of Millimeter-Wave Spectroscopy* (eds Latter, W. B., Radford, S. J. E., Jewell, P. R., Magnum, J. G. & Bally, J.) 108–111 (IAU Symp. Vol. 170, Kluwer Academic, Dordrecht, 1996).
 5. Kahane, C., Audinos, P., Barnbaum, C. & Morris, M. V Hydrae: the missing link between spherical red giants and bipolar planetary nebulae? Radio observations of the molecular envelope. *Astron. Astrophys.* **314**, 871–882 (1996).
 6. Knapp, G. R., Jorissen, A. & Young, K. A 200 km/sec molecular wind in the carbon star V Hya. *Astron. Astrophys.* **326**, 318–328 (1997).
 7. Lloyd Evans, T. Optical spectroscopy of the jet in V Hydrae. *Mon. Not. R. Astron. Soc.* **248**, 479–482 (1991).
 8. Knapp, G. R. *et al.* The light curve and evolutionary status of the carbon star V Hya. *Astron. Astrophys.* **351**, 97–102 (1999).
 9. Barnbaum, C., Morris, M. & Kahane, C. Evidence for rapid rotation of the carbon star V Hydrae. *Astrophys. J.* **450**, 862–875 (1995).
 10. Lee, C.-F. & Sahai, R. Shaping proto-planetary and young planetary nebulae with collimated fast winds. *Astrophys. J.* **586**, 319–337 (2003).
 11. Lloyd Evans, T. in *The Carbon Star Phenomenon* (ed. Wing, R. E.) Vol. 177, 367–376 (IAU Symp., Kluwer, Dordrecht, 2000).
 12. Hartigan, P., Raymond, J. & Hartmann, L. Radiative bow shock models of Herbig-Haro objects. *Astrophys. J.* **316**, 323–348 (1987).
 13. Livio, M. in *Accretion Phenomena and Related Outflows* (eds Wickramasinghe, D. T., Bicknell, G. V. & Ferrario, L.) 845–860 (ASP Conf. Ser. 121, IAU Colloq. 163, Astronomical Society of the Pacific, San Francisco, 1997).
 14. Morris, M. Mechanisms for mass loss from cool stars. *Publ. Astron. Soc. Pacif.* **99**, 1115–1122 (1987).
 15. Soker, N. & Rappaport, S. The formation of very narrow waist bipolar planetary nebulae. *Astrophys. J.* **538**, 241–259 (2000).
 16. Soker, N. Jet formation in the transition from the asymptotic giant branch to planetary nebulae. *Astrophys. J.* **389**, 628–634 (1992).
 17. Alcolea, J., Bujarrabal, V., Sánchez Contreras, C., Neri, R. & Zweigle, J. The highly collimated bipolar outflow of OH 231.8+4.2. *Astron. Astrophys.* **373**, 932–949 (2001).
 18. Imai, H., Obara, K., Diamond, P. J., Omodaka, T. & Sasao, T. A. A collimated jet of molecular gas from a star on the asymptotic giant branch. *Nature* **417**, 829–831 (2002).
 19. Kastner, J. H. *et al.* Compact X-ray source and possible X-ray jets within the planetary nebula Menzel 3. *Astrophys. J.* **591**, L37–L40 (2003).
 20. Kellogg, E., Pedelty, J. A. & Lyon, R. G. The X-ray system R Aquarii: A two-sided jet and central source. *Astrophys. J.* **563**, L151–L155 (2001).
 21. Bally, J. in *Accretion Phenomena and Related Outflows* (eds Wickramasinghe, D. T., Bicknell, G. V. & Ferrario, L.) 3–13 (ASP Conf. Ser. 121, IAU Colloq. 163, Astronomical Society of the Pacific, San Francisco, 1997).
 22. Reipurth, B., Heathcote, S., Morse, J., Hartigan, P. & Bally, J. Hubble Space Telescope images of the HH 34 jet and bow shock: Structure and proper motions. *Astron. J.* **123**, 362–381 (2002).
 23. Soker, N. Pairs of bubbles in planetary nebulae and clusters of galaxies. *Publ. Astron. Soc. Pacif.* **115**, 1296–1300 (2003).

Acknowledgements We thank H. Imai for commenting on an earlier version of this paper. We are grateful to NASA for partial financial support, obtained through: the Long Term Space Astrophysics programme for R.S. and M.M.; and grants for R.S. from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under a NASA contract.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.S. (raghvendra.sahai@jpl.nasa.gov).

Demonstration of an all-optical quantum controlled-NOT gate

J. L. O'Brien¹*, G. J. Pryde¹*, A. G. White¹, T. C. Ralph¹ & D. Branning^{1,2}

¹Centre for Quantum Computer Technology, Department of Physics, University of Queensland, Brisbane 4072, Australia

²Department of Physics, University of Illinois at Urbana-Champaign, Urbana Illinois 61801-3080, USA

* These authors contributed equally to this work

The promise of tremendous computational power, coupled with the development of robust error-correcting schemes¹, has fuelled extensive efforts² to build a quantum computer. The requirements for realizing such a device are confounding: scalable quantum bits (two-level quantum systems, or qubits) that can

be well isolated from the environment, but also initialized, measured and made to undergo controllable interactions to implement a universal set of quantum logic gates³. The usual set consists of single qubit rotations and a controlled-NOT (CNOT) gate, which flips the state of a target qubit conditional on the control qubit being in the state 1. Here we report an unambiguous experimental demonstration and comprehensive characterization of quantum CNOT operation in an optical system. We produce all four entangled Bell states as a function of only the input qubits' logical values, for a single operating condition of the gate. The gate is probabilistic (the qubits are destroyed upon failure), but with the addition of linear optical quantum non-demolition measurements, it is equivalent to the CNOT gate required for scalable all-optical quantum computation⁴.

Nuclear magnetic resonance techniques have been used to implement the most advanced quantum algorithms to date⁵. As they encode in mixed states and make ensemble measurements, these systems are not ultimately scalable. In contrast, ion trap systems have been used to implement high-fidelity two qubit quantum gates on pure states of trapped ions^{6,7}. Solid-state systems, including spin qubits in semiconductors⁸ and quantum dots², have been hailed for their potential scalability, and recently, entanglement between two superconducting qubits was demonstrated⁹. Single photon qubits offer the dual advantages of excellent isolation from the environment and ease of manipulation at the single qubit level, and have consequently found wide application in quantum cryptography protocols¹⁰. Photon qubits have the added advantage that an unequalled level of control over their quantum state makes comprehensive characterization possible, as demonstrated in the state tomography measurements presented here.

The difficulty in optical quantum computing has been in achieving the two photon interactions required for a two qubit gate (although progress has been made in cavity quantum electrodynamics systems¹¹). Knill, Laflamme and Milburn (KLM) have proposed a solution: a non-deterministic CNOT gate, where the required nonlinearity is accomplished using extra 'ancilla' photons—photons that are not part of the computation—and single photon detection. This gate can be made efficiently deterministic (scalable) by a teleportation protocol¹². A related proposal requires triggered entangled pairs of photons as a resource¹³. Other schemes show some, but not all, of the features of a quantum CNOT gate^{14,15} in that they cannot work for an arbitrary input state. The ultimate realization of the KLM CNOT gate will require: heralded single photon sources with stringent mode and bandwidth characteristics; high-efficiency number resolving single photon detectors; and construction of complicated optical circuits exhibiting both classical and quantum interference effects. Progress has been made towards reaching the first two requirements^{16–20}, and here we address the last by demonstrating the CNOT gate shown conceptually in Fig. 1a^{21,22}. Combined with quantum non-demolition (QND) measurement of the outputs, this gate is equivalent to the KLM CNOT: QND measurements can be made with additional single photon inputs, linear optics and number resolving single photon detection²³. In Fig. 1a the control (C) and target (T) qubits act as their own ancilla: the gate operates correctly conditional on simultaneous detection of a single photon in each of the outputs, which is assumed in the following discussion.

The gate works as follows²¹: the control and target qubits are each encoded by a single photon across two spatial modes—spatial encoding. As indicated in Fig. 1b, any arbitrary superposition state is possible: $\alpha|0\rangle + \beta|1\rangle$, where $|0\rangle$ and $|1\rangle$ are the two spatial modes corresponding to the logical basis states. The coefficients are normalized complex amplitudes: $|\alpha|^2 + |\beta|^2 = 1$. The two target modes are mixed and recombined on two 50% reflective beam splitters ($\frac{1}{2}$ BSs) to form an interferometer that also includes a 33% reflective beam splitter ($\frac{1}{3}$ BS) in each arm. This interferometer is

balanced so that in the absence of a control photon, the target qubit leaves in the same state that it entered. For the control in the state $|0\rangle$ this remains true because there is no interaction between the qubits. However, if the control is in the state $|1\rangle$ the control and target photons interfere non-classically at the central $\frac{1}{3}$ BS owing to path indistinguishability²⁴. This two-photon quantum interference causes a π phase shift in the upper arm of the target interferometer and results in the target state output being flipped: $\alpha|0\rangle + \beta|1\rangle \rightarrow \beta|0\rangle + \alpha|1\rangle$. The control qubit's logical value is unchanged. Because of the $\frac{1}{3}$ BSs we do not always observe a single photon in each of the control and target outputs. However, when we do detect a single photon in each output (a coincidence count), which occurs with probability $P = \frac{1}{9}$, we know that the CNOT operation has been correctly realized. The most important feature of a CNOT gate is its quantum operation: with the control in a superposition and the target in a logical basis state, the output of the CNOT gate is an entangled state—the quintessential quantum mechanical state, necessary for universal quantum computation.

It is most practical to prepare single photon qubits where the quantum information is encoded in the polarization state $\alpha|H\rangle + \beta|V\rangle$ ($\equiv \alpha|0\rangle + \beta|1\rangle$)—polarization encoding, where $|H\rangle$ and $|V\rangle$ are the horizontal and vertical polarization states, respectively. To convert to spatial encoding requires a polarizing beam splitter (PBS) and half-wave plate (HWP) (Fig. 1b). To convert from polarization, to spatial and back to polarization encoding, while preserving the quantum information, requires that the phase relationship between the two basis components be preserved throughout: the path lengths must be subwavelength stable (interferometric stability). Therefore, the CNOT gate shown schematically in Fig. 1a requires two classical interferences and one non-classical interference of the

control and target photons, to be satisfied simultaneously—a significant challenge.

To meet these requirements we have designed the inherently stable interferometer shown in Fig. 1c, where there is a one-to-one mapping from the conceptual schematic of Fig. 1a. The target interferometer is realized by mixing and recombining the logical basis modes while the target qubit is polarization encoded: this requires a HWP set to rotate the polarization by 45° (that is, with its optic axis $OA = 22.5^\circ$), which equally mixes the two polarization modes (a Hadamard gate (H) in the language of quantum information: the CNOT gate can be thought of as a controlled-phase shift gate with a Hadamard gate at the input and output of the target). Transformation to spatial encoding occurs at a PBS where the two output modes are parallel but displaced. The operation of all three $\frac{1}{3}$ BSs is realized by a single $\frac{1}{3}$ HWP ($OA = 62.5^\circ$): the C_1 and T_+ components are orthogonally polarized and leave the first PBS in the same spatial mode. They are unequally mixed on the $\frac{1}{3}$ HWP and the required quantum interference is realized as the photons are only partly distinguishable after the second PBS. The OA setting is chosen to achieve this interference and also to swap the roles of H and V so that the control and target modes recombine correctly at the second PBS. In addition, the C_0 and T_- modes are polarization-rotated by the $\frac{1}{3}$ HWP such that, after the second PBS, $\frac{2}{3}$ of each component goes into the dumps, as required to balance the gate. The total two-qubit output state of the gate is analysed via quantum state tomography.

To characterize the operation of this gate we first measured the output of the gate for each of the four possible logical basis input states: $|C\rangle|T\rangle = |CT\rangle = |00\rangle, |01\rangle, |10\rangle$ and $|11\rangle$. The correct behaviour is represented in Fig. 2a and compared to that observed

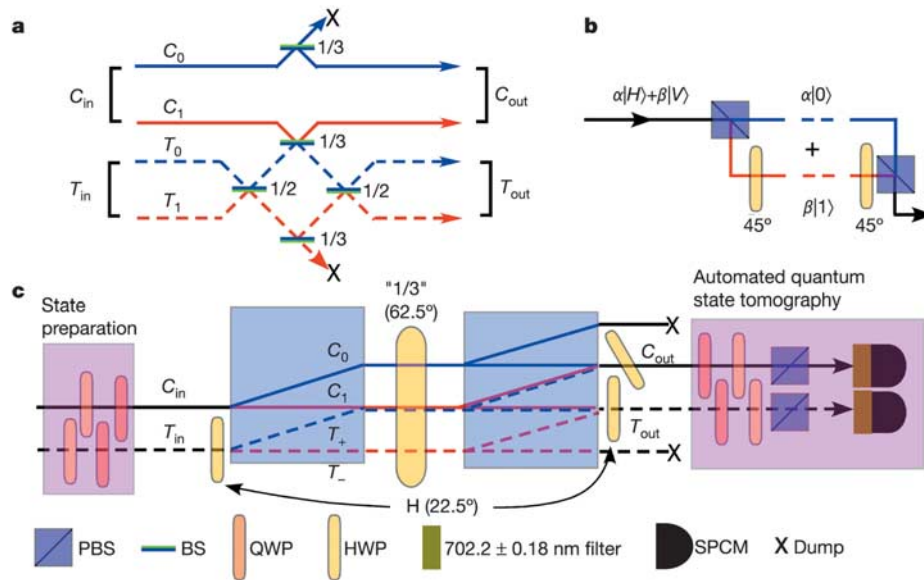


Figure 1 A schematic of the CNOT gate realized in this work. **a**, A conceptual depiction of the gate, as described in the text. A sign change (π phase shift) occurs upon reflection off the green side of the beam splitters (BSs). **b**, A polarization encoded photonic qubit can be converted into a spatially encoded qubit, suitable for the gate shown in **a**, using a polarizing beam splitter (PBS) and a half-wave plate (HWP) set to rotate the polarization of one of the outputs by 90° ($OA = 45^\circ$). The rotation is required so that all components of the spatial qubits have the same polarization and can interfere both classically and non-classically. The reverse process converts the spatial encoding back to polarization encoding. **c**, A schematic of the experimental CNOT gate. Pairs of energy degenerate photons are incident from the left of the diagram. These were generated through beam-like spontaneous parametric downconversion³⁰ and collected into single-mode optical fibres²⁵ (not shown). The output of each fibre is collimated and a HWP and a quarter-wave

plate (QWP) in each input beam allows preparation of any pure, separable two qubit state to be input into the gate. The horizontal and vertical components of the qubits are separated and recombined using PBSs made from the birefringent material calcite, where the output modes are parallel and displaced. This interferometer is inherently stable, being insensitive to translation of the PBSs. The two outputs are polarization analysed using an automated tomography system consisting of a computer-controlled HWP and QWP followed by a PBS in front of each single photon counting module (SPCM). Simultaneous detection of a single photon at each of the detectors—a coincidence count—signals that the gate has worked. A coincidence window of 5 ns was used throughout. The tilted HWP at 0° is fixed to correct a phase shift in the control interferometer.

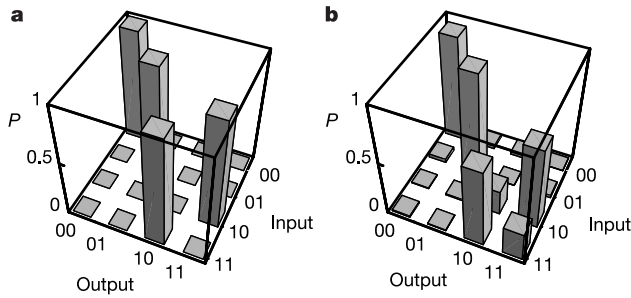


Figure 2 Experimental demonstration of classical CNOT operation—operation in the logical basis. **a**, Ideal logical basis operation of a CNOT gate. **b**, Measured operation for the gate presented here.

experimentally (Fig. 2b). The gate works very well for the $|C\rangle = |0\rangle$ inputs ($94 \pm 2\%$ and $95 \pm 2\%$), which reflects the fact that only a single classical interference is required for correct operation. For the $|C\rangle = |1\rangle$ inputs the gate works less well ($75 \pm 2\%$ and $72 \pm 2\%$), owing to the added non-classical interference required (Table 1). The probability of getting the correct output averaged over all logical inputs is 84%. This value compares favourably with that obtained from the data of ref. 6: 73.5%. However, logical basis operation is purely classical and therefore demonstrates only part of the required action of a quantum CNOT gate.

The next step is to demonstrate that the gate is entangling—it produces an entangled two qubit output state from a separable input—which can be done by measuring conditional fringe visibilities²⁵. Figure 3 shows plots of coincident photon count rates as a function of the HWP setting in the target analyser. The two curves are for the control analyser HWP set to analyse $|0\rangle$ and $|0\rangle + |1\rangle$, respectively. In both cases the input state is $(|0\rangle - |1\rangle)_C |1\rangle_T$, which ideally produces the maximally entangled Bell singlet state $|\Psi^-\rangle = |01\rangle - |10\rangle$. The visibilities ($v = (\max - \min)/(\max + \min)$) for the fitted curves, for which the period is fixed at 90° , are $93.5 \pm 2\%$ and $92 \pm 3\%$ respectively. High-visibility fringes in non-orthogonal bases like these are a classic signature of entanglement²⁵, providing evidence for the quantum operation of the gate. However, they do not provide enough information to reconstruct the output state because the degree of mixture is unknown.

Complete state reconstruction is possible using quantum state tomography: a series of measurements on a large number of identically prepared copies of a quantum system allows accurate estimation of the quantum state of that system. In the case of two qubits, this requires 16 different joint measurements of the two qubit state²⁶, which can be used to reconstruct the density matrix, which contains everything that can be known about the two qubit state. We propose that production and quantum state tomography of the four maximally entangled Bell states with high fidelity is an important functional demonstration for a CNOT gate. Figure 4a shows the density matrix of $|\Psi^-\rangle$, and the real and imaginary parts of the reconstructed density matrix $\hat{\rho}$ of the output state of our gate for the input $(|0\rangle - |1\rangle)_C |1\rangle_T$. The fidelity is $F_{\Psi^-} = \langle \Psi^- | \hat{\rho} | \Psi^- \rangle = 0.87 \pm 0.08$. Also shown (Fig. 4b) are the density matrix for $|\Phi^+\rangle$ and the real and imaginary components of the reconstructed density matrix for the input state $(|0\rangle + |1\rangle)_C |0\rangle_T$, where $F_{\Phi^+} = 0.77 \pm$

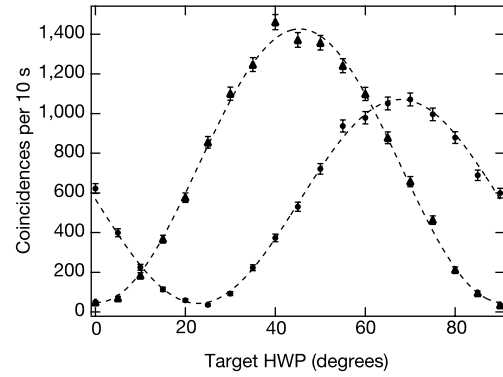


Figure 3 Conditional coincidence fringes for non-orthogonal bases. The control analyser was set to pass $|0\rangle + |1\rangle$ (circles) and $|0\rangle$ (triangles) for the input state $(|0\rangle - |1\rangle)_C |1\rangle_T$. The error bars corresponding to HWP angle are too small to see on this plot. The fitted curves have a period that is fixed at 90° and the phase offsets are $0.4 \pm 0.2^\circ$ and $0.3 \pm 0.2^\circ$, respectively.

0.09. The fidelities for the other two Bell states (not plotted) are shown in Table 2, along with measures of entanglement and mixture for all four Bell states. In all cases, the experimentally measured output state is in the range where a Bell inequality can be violated²⁷.

All of the data shown here were taken over one day for a single operating condition of the gate, and its performance was observed to be repeatable. This demonstrates two important points: the gate is very stable, and it does not require tuning for different input states. In addition, no correction has been made for accidental coincidence counts, which will introduce small errors in gate operation. By far the largest source of error in our gate is due to decoherence, which arises from imperfect mode matching for the non-classical interference. This can be seen clearly in the logical basis operation shown in Fig. 2: the gate works less well for states where the control is in the logical $|1\rangle$ state and the error terms are due to mode mismatch between the C_1 and T_+ modes, resulting in the target not being flipped as required. The errors in production of the $|\Phi^+\rangle$ state are also due to mismatch of these modes: in the experimentally reconstructed density matrix (Fig. 4b) we can see that the error terms are residual components of the input state, with the original coherences preserved. The differences between the fidelities of the four Bell states—which all require non-classical interference because $|C\rangle = |0\rangle \pm |1\rangle$ —is understood to arise from small amounts of input beam steering introduced by the state preparation waveplates.

We have demonstrated a two photon CNOT gate operating via coincident photon detection. In the logical basis the gate operates with an average success of 84%. Conditional fringe visibilities exceeding 90% in non-orthogonal bases indicate entanglement. Complete quantum state tomography confirms this, showing production of all four entangled Bell states with fidelities greater than 75%—an important functional demonstration of quantum CNOT operation. Note that these results go well beyond a recent report of an alternative optical CNOT gate²⁸ that shows logical basis opera-

Table 1 Experimentally determined probabilities for the logical basis operation

Input $ CT\rangle$	$P_{ 00\rangle}$	$P_{ 01\rangle}$	$P_{ 10\rangle}$	$P_{ 11\rangle}$
$ 00\rangle$	0.95(2)	0.023(3)	0.024(3)	0.0006(5)
$ 01\rangle$	0.031(3)	0.94(2)	0.0019(8)	0.022(3)
$ 10\rangle$	0.005(1)	0.011(2)	0.23(9)	0.75(2)
$ 11\rangle$	0.011(2)	0.0005(1)	0.72(2)	0.26(1)

Data as plotted in Fig. 2.

Table 2 Characterization of the four Bell states

State	Fidelity	Tangle	Linear entropy
$\Psi^- = 01\rangle - 10\rangle$	0.87(8)	0.65(6)	0.27(6)
$\Psi^+ = 01\rangle + 10\rangle$	0.75(9)	0.55(7)	0.31(5)
$\Phi^- = 00\rangle - 11\rangle$	0.76(9)	0.46(5)	0.45(3)
$\Phi^+ = 00\rangle + 11\rangle$	0.77(9)	0.49(9)	0.45(5)

The degree of entanglement of any state can be measured by calculating the tangle $T = C^2$, where C is the concurrence^{26,27}. Similarly, the degree of mixture can be measured by calculating the linear entropy^{26,27}. These values, along with fidelities, are given for all four experimentally produced Bell states.

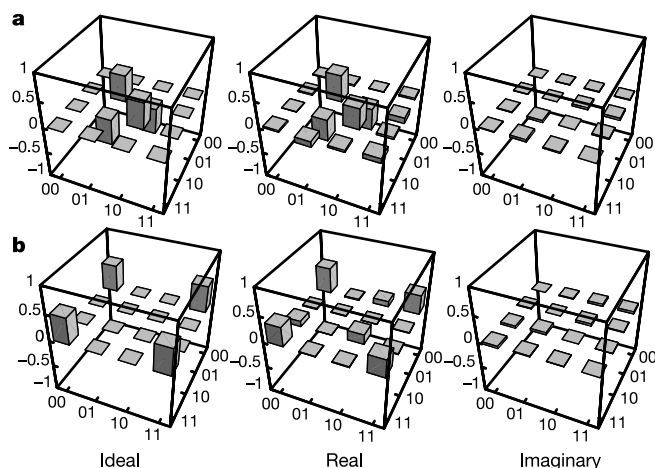


Figure 4 Density matrices for two highly entangled output states. **a**, The real part of the density matrix for the maximally entangled Bell singlet state $|\Psi^-\rangle = |01\rangle - |10\rangle$ (the imaginary components are all zero), and the real and imaginary parts of the density matrix reconstructed from quantum process tomography for the input state $(|0\rangle - |1\rangle)_d|1\rangle_T$. **b**, The real part of the $|\Phi^+\rangle$ state, and the real and imaginary parts of the density matrix reconstructed from quantum process tomography for the input state $(|0\rangle + |1\rangle)_d|0\rangle_T$.

tion, but only a single coincidence fringe with $\nu = 61.5 \pm 7.4\%$, making the issue of entanglement ambiguous. The CNOT gate presented here combined with QND is equivalent to the KLM CNOT gate, and could be made scalable with the same teleportation protocol. The next step will be to incorporate the gate reported here in simple optical circuits to demonstrate simple algorithms and error correcting schemes²⁹. □

Received 26 June; accepted 16 September 2003; doi:10.1038/nature02054.

- Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, 2000).
- Clark, R. G. (ed.) *Quant. Inform. Comput.* 1 (special issue on implementation of quantum computation) 1–50 (2001).
- DiVincenzo, D. P. & Loss, D. Quantum information is physical. *Superlatt. Microstruct.* 23, 419–432 (1998).
- Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* 409, 46–52 (2001).
- Vandersypen, L. M. K. *et al.* Experimental realisation of Shor's quantum factoring algorithm using nuclear magnetic resonance. *Nature* 414, 883–887 (2001).
- Schmidt-Kaler, F. *et al.* Realisation of the Cirac–Zoller controlled-NOT quantum gate. *Nature* 422, 408–411 (2003).
- Leibfried, D. *et al.* Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate. *Nature* 422, 412–415 (2003).
- Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* 393, 133–137 (1998).
- Pashkin, Yu. A. *et al.* Quantum oscillations in two coupled charge qubits. *Nature* 421, 823–826 (2003).
- Gisin, N., Ribordy, G., Tittel, W. & Zbinden, H. Quantum cryptography. *Rev. Mod. Phys.* 74, 145–195 (2002).
- Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H. & Kimble, H. J. Measurement of conditional phase shifts for quantum logic. *Phys. Rev. Lett.* 75, 4710–4713 (1995).
- Gottesman, D. & Chuang, I. L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* 402, 390–393 (1999).
- Koashi, M., Yamamoto, T. & Imoto, N. Probabilistic manipulation of entangled photons. *Phys. Rev. A* 63, 030301 (2001).
- Pan, J.-W., Simon, C., Brukner, Č. & Zeilinger, A. Entanglement purification for quantum communication. *Nature* 410, 1067–1070 (2001).
- Pittman, T. B., Jacobs, B. C. & Franson, J. D. Probabilistic quantum logic operations using polarizing beam splitters. *Phys. Rev. A* 64, 062311 (2001).
- Santori, C., Fattal, D., Vučković, J., Solomon, G. S. & Yamamoto, Y. Indistinguishable photons from a single-photon device. *Nature* 419, 594–597 (2002).
- Kuhn, A., Hennrich, M. & Rempe, G. Deterministic single-photon source for distributed quantum networking. *Phys. Rev. Lett.* 89, 067901 (2002).
- Kuzmich, A. *et al.* Generation of nonclassical photon pairs for scalable quantum communication with atomic ensembles. *Nature* 423, 731–734 (2003).
- James, D. F. V. & Kwiat, P. G. Atomic-vapor-based high efficiency optical detectors with photon number resolution. *Phys. Rev. Lett.* 89, 183601 (2002).
- Imamoglu, A. High efficiency photon counting using stored light. *Phys. Rev. Lett.* 89, 163602 (2002).
- Ralph, T. C., Langford, N. K., Bell, T. B. & White, A. G. Linear optical controlled-NOT gate in the coincidence basis. *Phys. Rev. A* 65, 062324 (2001).
- Hofmann, H. F. & Takeuchi, S. Quantum phase gate for photonic qubits using only beam splitters and postselection. *Phys. Rev. A* 66, 024308 (2001).

- Kok, P., Lee, H. & Dowling, J. P. Single-photon quantum-nondemolition detectors constructed with linear optics and projective measurements. *Phys. Rev. A* 66, 063814 (2002).
- Hong, C. K., Ou, Z. Y. & Mandel, L. Measurement of subpicosecond time intervals between two photons by interference. *Phys. Rev. Lett.* 59, 2044–2046 (1987).
- Kurtsiefer, C., Oberparleiter, M. & Weinfurter, H. High-efficiency entangled photon pair collection in type-II parametric fluorescence. *Phys. Rev. A* 64, 023802 (2001).
- James, D. F. V., Kwiat, P. G., Munro, W. J. & White, A. G. Measurement of qubits. *Phys. Rev. A* 64, 052312 (2001).
- Munro, W. J., Nemoto, K. & White, A. G. The Bell inequality: A measure of entanglement? *J. Mod. Opt.* 48, 1239–1246 (2001).
- Pittman, T. B., Fitch, M. J., Jacobs, B. C. & Franson, J. D. Experimental controlled-NOT logic gate for single photons. Preprint at (<http://arXiv.org/quant-ph/0303095>) (2003).
- Dodd, J. L., Ralph, T. C. & Milburn, G. J. Experimental requirements for Grover's algorithm in optical quantum computation. *Phys. Rev. A* (in the press); preprint at (<http://arXiv.org/quant-ph/0306081>) (2003).
- Takeuchi, S. Beamlike twin-photon generation by use of type II parametric downconversion. *Opt. Lett.* 26, 843–845 (2001).

Acknowledgements We thank N. K. Langford for experimental work related to non-classical interference, T. B. Bell for work on the quantum state tomography system, and P. T. Cochrane, J. L. Dodd, A. Gilchrist, P. G. Kwiat, G. J. Milburn, W. J. Munro and M. A. Nielsen for discussions. This work was supported by the Australian government, the Australian Research Council, the US National Security Agency (NSA) and Advanced Research and Development Activity (ARDA) under the Army Research Office (ARO).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.L.O'B. (job@physics.uq.edu.au).

Direct observation of attosecond light bunching

P. Tzallas¹, D. Charalambidis², N. A. Papadogiannis², K. Witte¹ & G. D. Tsakiris¹

¹Max-Planck-Institut für Quantenoptik, D-85748 Garching, Germany

²Foundation for Research and Technology-Hellas, Institute of Electronic Structure & Laser, PO Box 1527, GR-711 10 Heraklion (Crete), and Department of Physics, University of Crete, PO Box 2208, GR-71003 Voutes-Heraklion (Crete), Greece

Temporal probing of a number of fundamental dynamical processes requires intense pulses at femtosecond or even attosecond ($1 \text{ as} = 10^{-18} \text{ s}$) timescales. A frequency 'comb' of extreme-ultraviolet odd harmonics can easily be generated in the interaction of subpicosecond laser pulses with rare gases: if the spectral components within this comb possess an appropriate phase relationship to one another, their Fourier synthesis results in an attosecond pulse train^{1,2}. Laser pulses spanning many optical cycles have been used for the production of such light bunching^{3,4}, but in the limit of few-cycle pulses the same process produces isolated attosecond bursts^{5,6}. If these bursts are intense enough to induce a nonlinear process in a target system^{7–9}, they can be used for subfemtosecond pump–probe studies of ultrafast processes. To date, all methods for the quantitative investigation of attosecond light localization^{4,6,10} and ultrafast dynamics¹¹ rely on modelling of the cross-correlation process between the extreme-ultraviolet pulses and the fundamental laser field used in their generation. Here we report the direct determination of the temporal characteristics of pulses in the subfemtosecond regime, by measuring the second-order autocorrelation trace of a train of attosecond pulses. The method exhibits distinct capabilities for the characterization and utilization of attosecond pulses for a host of applications in attoscience.

This emerging field is a natural follow-up to the study of a multitude of ultrafast dynamical systems based on the tools provided by progress in femtosecond laser pulse engineering¹². How-

ever, several other natural processes—such as electron–electron correlation effects in all states of matter, or the early stages of energy redistribution in antenna complexes within the photosynthetic reaction centre—evolve in characteristic times of a few femtoseconds or less. Obtaining ‘snapshots’ of their evolution with subfemtosecond temporal resolution is the goal that the recently initiated and rapidly developing attoscience strives for. Indeed, only ten years ago was it first suggested that extreme-ultraviolet (XUV) harmonic emission from rare gases might be a suitable candidate for the temporal localization of light to unprecedented short timescales in the form of an attosecond pulse train². More recently, conditions for the generation of isolated attosecond pulses using few-cycle light pulses have also been established⁵. The attosecond-scale dynamics of the nonlinearly driven atomic electron wave packets that generate the XUV radiation have been outlined in theoretical studies^{13–15}, while a number of experiments have revealed complexities related to these dynamics^{16–18} and others have provided evidence that coherent control of these processes is possible and advantageous^{19–20}.

In the quest for the generation of attosecond pulses, both the many- and the few-cycle pulse routes are being at present pursued^{3,4,6}. Although attosecond structures have previously been reported, all methods used so far for their investigation^{4,6,10} are (owing to the lack of nonlinear XUV autocorrelators) based on sophisticated cross-correlation approaches with the fundamental laser field. Although they have decisively contributed to the development of attoscience in its early stage, the interpretation of the data and their use in extracting quantitative results inevitably requires modelling based on appropriate approximations and/or use of theoretically calculated parameters. The recent debut of attosecond pulses in pump–probe applications¹¹ relies also on such cross-correlation techniques. In this case, the accuracy with which the laser field is known and that of the model used determine the temporal resolution of the technique. As attoscience matures, direct high-resolution time-domain metrology approaches become more important.

Here we present an essential step forward in gaining access to attosecond timescales. It is based on nonlinearly autocorrelating XUV radiation through its interaction with an atomic system—that is, on a non-trivial extension of a well established approach used successfully for many years in optical femtosecond metrology. Short-wavelength second-order autocorrelation has been limited so far to measuring the duration of individual XUV⁷ or UV²¹ harmonics in the temporal regime of a few tens of femtoseconds. The present work is important in expanding the applicability of this nonlinear technique to the broadband radiation of a coherent superposition of harmonics and thus to the attosecond regime. Although the unique capabilities of the new approach are demonstrated here in performing the first autocorrelation measurement of an attosecond pulse train, the method itself initiates XUV-pump–XUV-probe studies in subfemtosecond-scale dynamics. In principle, it is a model-independent method that provides a direct temporal resolution of the order of the width of the attosecond wave packet. Apart from its importance to the characterization, optimization and time-domain applications of harmonic attosecond fields, the method becomes highly pertinent in connection with novel nonlinear pump–probe experiments using the XUV free electron laser sources that have recently opened up a new field of nonlinear physics²².

The mapping of phase-amplitude distributions of an ultrashort pulse or an ultrafast process relies on a nonlinear process. In femtosecond metrology, the most widely used methods rely on a nonlinear effect induced solely by the radiation to be characterized. In pico- and femtosecond laser laboratories, the pulse duration has been for many years routinely extracted to a satisfactory degree of accuracy from a measurement of the second-order autocorrelation trace. The extension of the approach to subfemtosecond XUV pulses poses several formidable problems, because attosecond

pulses are spectrally much broader and in the nearly inaccessible UV–XUV spectral range, and are orders of magnitude weaker, thus requiring ultrasensitive nonlinear detectors with a flat broadband response. The key factors in surmounting these obstacles are (1) the technique of the dispersionless volume autocorrelation and (2) the demonstration of the two-photon non-resonant ionization process in helium as an appropriate detector with adequately flat response. The experimental results bear direct evidence that a synthesis of five harmonics give rise to an XUV signal exhibiting clear attosecond structure with periodicity twice that of the driving laser field. An approximate analysis of the signal gives ~ 780 as average duration for the attosecond peaks.

In the experiment, the harmonic generation takes place in a xenon gas jet using 130-fs laser pulses at wavelength $\lambda = 790$ nm of

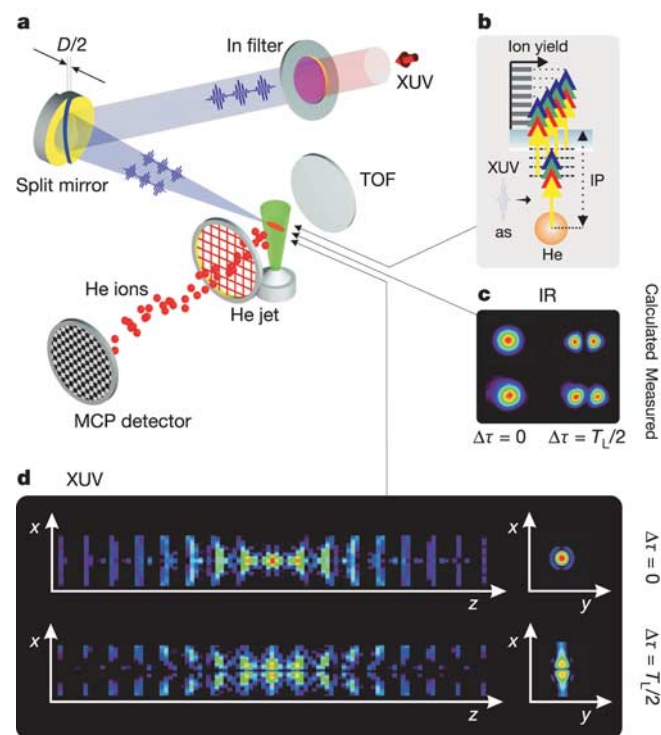


Figure 1 The second-order XUV autocorrelator. **a**, Schematic representation of the set-up. Only the seventh to the fifteenth harmonics of the XUV radiation generated at the Xe jet (not shown) are let through by the In filter and enter the volume autocorrelator. The second-order XUV autocorrelator consists of a spherical mirror, split into two halves, serving as a focusing wavefront divider. As a nonlinear detector, a two-photon ionized He gas is used and the He ion yield, recorded by a time-of-flight (TOF) mass spectrometer, provides the autocorrelation signal. The ionization occurs through two-XUV-photon non-resonant absorption from all possible combinations of the transmitted harmonics, as shown schematically in **b** for the case of the ninth to the fifteenth harmonics. Panels **c** and **d** show the principle of the volume autocorrelator. **c**, For zero and $\Delta\tau = T_L/2$ delay, the calculated and measured transverse intensity distribution for the laser frequency (IR) changes from an Airy spot to a double maximum distribution. The three-dimensional colour contours in **d** depict, in logarithmic scale over three decades and for a displacement of 0 and $D = \lambda/2$ ($\Delta\tau = T_L/2$ delay), the simulated XUV intensity distribution $I(D, x, y, z) = E(D, x, y, z)E^*(D, x, y, z)$ at the ionization region, where $E(D, x, y, z) = \sum_{n=7}^{15} E_n(D, x, y, z)$ is the total field of the superposition of the seventh to the thirteenth harmonics having equal amplitudes and phases. The pattern extends over a box with transverse sides (x - and y -direction) six times the Airy spot radius, and length (z -direction) six times the Rayleigh range of the focusing mirror, for the wavelength of the ninth harmonic. Its complexity is a consequence of spatial beating due to the multiplicity of the wavelengths present, but as in the case of single wavelength (**c**), for a delay of $T_L/2$ the spatial distribution is split into two parts. Although the total energy remains the same, this variation in the spatial distribution allows for second-order autocorrelation measurements.

up to 10-mJ energy from a 10-Hz Ti:sapphire laser. To reduce the amount of the fundamental after the xenon jet, an annular laser beam²³ is used. The laser beam focus was ~ 6 mm before the Xe gas jet; at this position, theoretical calculations predict optimum harmonic phase locking²⁴. For the second-order autocorrelation of this harmonic superposition, we used a wavefront splitting arrangement^{25,26} consisting of a spherical mirror with 30-cm radius of curvature cut into two halves (Fig. 1a). The positioning of one of these halves is controlled by a piezo-crystal translation unit with a resolution of ~ 6 nm. The two parts of the bisected XUV pulse train are brought into a common focus in a helium gas jet. The ionization products are detected by a time-of-flight mass spectrometer as a function of the delay $\Delta\tau$ corresponding to a total displacement D (twice the actual separation) between the two half mirrors. For the laser intensities used, harmonics up to the fifteenth are generated. A 0.2- μm indium filter selects a group of harmonics from the seventh to the fifteenth, and blocks the residual fundamental. The recorded harmonic spectrum is shown before (Fig. 2a) and after (Fig. 2b) the In filter. The corrected (Fig. 2c) relative intensity amplitude is 7:9:11:13:15 $\rightarrow$ 0.32:1.0:0.30:0.11:0.01. An ideal phase-locking of these harmonics (all phase differences equal to zero) would produce a train of attosecond pulses ('wagons') with full-width at half-maximum (FWHM) 'wagon' duration of 315 as.

Owing to the judicious choice of the combination of He as ionization medium and In as filter material, the harmonics in the transmitted superposition can induce only two-photon ionization, either by two photons of the same energy (with the exception of the seventh harmonic) or by any combination of the photon energies in the superposition as depicted in Fig. 1b. The properties of this nonlinear detector have been investigated previously⁷⁻⁹, and the spectral yields for the same harmonic superposition calculated from the time-dependent Schrödinger equation for He in a polychromatic laser field have revealed a flat detector response⁹. In order to verify the exclusive contribution of a two-photon ionization, we have performed measurements of the ion yield versus harmonic intensity for He⁺ and for the rest gas H₂O⁺, and Xe⁺ ions observed in the recorded mass spectra. Figure 2d-f gives the ion yield variation as a function of the ninth and eleventh harmonic intensities for the He⁺ ions, and of the eleventh harmonic for H₂O⁺ and Xe⁺ ions. The fact that the intensity dependence for He⁺ (ionization potential IP = 24.6 eV) is very nearly quadratic, while for H₂O⁺ (IP = 12.6 eV) and Xe⁺ (IP = 12.1 eV) it is linear, provides clear evidence that contributions to the He ionization due to processes other than the two-photon absorption are negligible. Additionally, the calculated ion yield for an estimated XUV intensity of $\sim 10^{11} \text{ W cm}^{-2}$ using the known He two-photon absorption cross-section^{8,9} is entirely consistent with the observed number of ~ 90 ions per pulse.

There are two points that distinguish the volume autocorrelation technique from the conventional Michelson interferometer based on a nonlinear crystal or a nonlinear photodiode. First, the signal produced emanates from the interaction of the XUV pulse with the He atoms within a volume defined by the focusing properties of the spherical mirror and not from a thin slice, as in the case of a nonlinear crystal. Second, unlike the amplitude splitting arrangements, for the wavefront splitting mirror device a delay variation results not in a change of the energy reaching the detector, but simply in a spatial redistribution of the energy in the focal volume. This is shown for the fundamental laser frequency in Fig. 1c, where a displacement of $D = \lambda/2$ (delay $\Delta\tau = T_L/2$, where T_L is the laser period of ~ 2.7 fs) between the two half mirrors has the effect of dividing the focal spot for zero displacement (Airy spot) with peak intensity I_{max} into two spots having approximately the same size as the original. As the total energy remains constant, the maximum intensity in each of the two spots is $I \approx I_{\text{max}}/2$. Using a nonlinear detector, the measured signal would undergo modulation because, for example, for a second-order process $S_{D=0} \propto I_{\text{max}}^2 \neq S_{D=\lambda/2} \propto$

$2(I_{\text{max}}/2)^2$. As shown in Fig. 1d, for a superposition of harmonics the spatial intensity pattern and its redistribution with delay becomes more complex. Our simulations indicate that a two-photon ionization signal $S(D) \propto \int \int \int I^2(D, x, y, z) dx dy dz$ would exhibit a modulation with delay and thus a second-order interferometric or intensity autocorrelation trace can be recorded. Furthermore, the resulting peak-to-background ratio is only $\sim 2.75:1$ for the volume autocorrelation trace instead of the 8:1 of the conventional interferometric autocorrelation trace. The corresponding ratio of an intensity autocorrelation is reduced from 3:1 (conventional) to $\sim 2:1$ (volume).

The results of the second-order autocorrelation measurement of a train of XUV pulses offering direct observation of attosecond temporal light confinement are summarized in Fig. 3. The middle panel gives the trace of the ion signal for a total temporal overlap near the peak of the laser pulse of ~ 18 fs in steps of ~ 37 as. For each position of the piezo-crystal, ten data points were accumulated (grey dots). Their average (yellow points) with the corresponding standard deviation evinces an attosecond pulse train. A distinct modulation comprising two peaks within an optical cycle of the fundamental frequency is clearly perceived. A coarser scan over ~ 100 fs has given an overall duration for the pulse train of ~ 60 fs. A

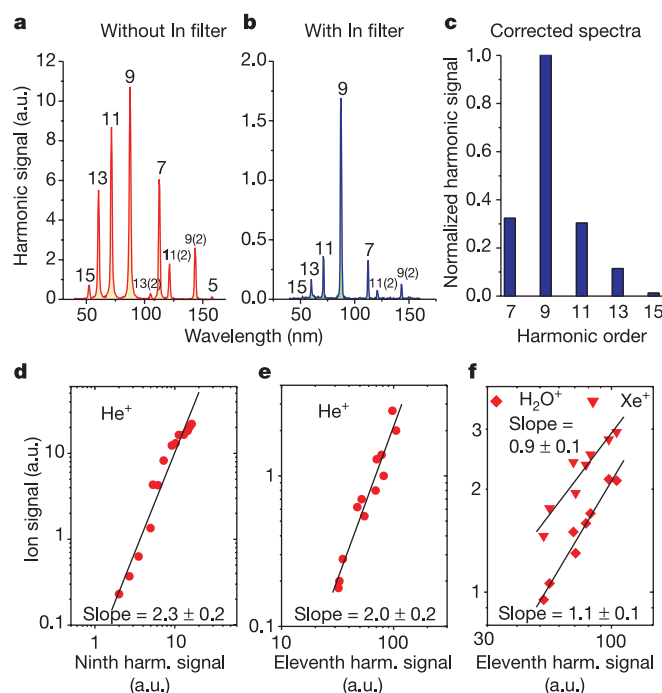


Figure 2 Ion yield dependence on the XUV radiation. **a–c**, Higher-order harmonic generation spectra produced in the Xe jet. **a**, As measured without the In filter; **b**, transmitted through the In filter; and **c**, after correction of the transmitted intensities for the reflectivity of the gold mirror and the efficiency of the XUV spectrometer and detector. In **c**, the relative intensities read 0.32 (seventh): 1.0 (ninth): 0.30 (eleventh): 0.11 (thirteenth): 0.01 (fifteenth). **d–f**, XUV intensity dependences of some of the masses. In log-log scale, the slope of the He ion-yield as a function of the intensity of the ninth harmonic (**d**) and the eleventh harmonic (**e**) is 2.3 ± 0.2 and 2.0 ± 0.2 respectively. This is in excellent agreement with a two-XUV-photon ionization process according to the lowest-order perturbation theory (LOPT) valid for the present experimental conditions. In contrast, the slope of the ion-yield dependence on the intensity of the eleventh harmonic for H₂O and Xe (**f**) is 0.9 ± 0.1 and 1.1 ± 0.1 respectively, again in perfect agreement with the LOPT for a single-photon ionization process. The difference in the order of the ionization process for the different species is due to their different ionization potentials: 24.6 eV, 12.1 eV and 12.6 eV for He, Xe and H₂O respectively. These results unambiguously prove a two-XUV-photon ionization of He on which the second-order autocorrelation measurements are based.

reliable *in situ* calibration of the optical period is afforded by recording the signal due to multiphoton ionization of He using laser light. This higher-order autocorrelation trace is shown in Fig. 3a.

An average duration over all the 'wagons' in the train can be inferred from our autocorrelation data by extending the standard procedure employed in second-order autocorrelation measurements of single pulses and performing fitting of the sum of a sequence of gaussian pulses with their common duration τ_{AC} as a single fit parameter. The red curve in Fig. 3b and c is the best fit obtained for $\tau_{AC} = 1.1$ fs. The average pulse duration of the 'wagon' in the attosecond train is thus $\tau_{XUV} = \tau_{AC}/\sqrt{2} = 780$ as. In the magnified section (Fig. 3c), we have plotted in addition the autocorrelation trace for $\tau_{XUV} = 700$ as and 860 as. The modulation is

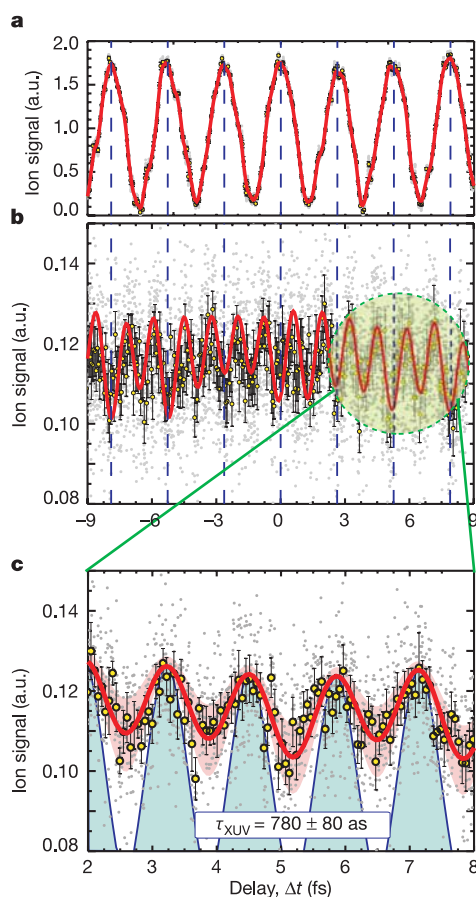


Figure 3 Measured nonlinear volume autocorrelation traces. **a**, The higher-order autocorrelation trace of the fundamental laser field. For this acquisition, the harmonic generation was turned off and the In filter was removed, thus ionization of He occurs through the fundamental laser frequency only. This trace is used for the calibration of the delay scale of the measured autocorrelation traces. The period of the observed oscillation is equal to the laser period, that is, 2.63 fs. **b**, A measured ~18-fs-long second-order intensity volume autocorrelation trace of the superposition of the five harmonics, seventh to fifteenth. The step of the autocorrelator is ~37 as, and 10 points are accumulated at each position (grey dots). The yellow points represent their average, and the error bars shown correspond to one standard deviation. A clear modulation with half the laser period is observable in the entire trace. The red curve is the best fit of an analytical function given by the sum of a sequence of gaussian pulses. The common temporal width, τ_{AC} , of each pulse is the fit parameter, with a best value being 1.1 fs, which corresponds to an average duration of $\tau_{XUV} = 780$ as for each XUV pulse in the train. **c**, An expanded area of the trace in **b**. The shaded blue lines are the individual (not superimposed) gaussian pulses with $\tau_{XUV} = 780$ as. As the light red shaded area indicates, the scatter in the experimental points gives an estimate of $\tau_{XUV} = 780 \pm 80$ as for average duration of the 'wagon' in the synthesis of the five harmonics.

accordingly enhanced or reduced so that it encompasses the breadth of the scatter in the data points, thus giving an estimate of $\tau_{XUV} = 780 \pm 80$, as for the average duration range of the attosecond peak. Although the second-order autocorrelation technique provides a direct duration measurement, the result depends on the pulse shape assumed and in cases of complex pulse shapes like those encountered in this experiment, the task of a more rigorous deconvolution becomes rather formidable. For example, no information can be ascertained as to whether the individual 'wagons' have variable duration and/or spacing between them. Nevertheless, within the accuracy limitations of the second-order autocorrelation technique, the simple deconvolution procedure that we have adopted here gives an average duration, which is more than double the Fourier-transform-limited value. This points to phase correlated spectral components but with imperfect phase locking of the harmonics. It also illustrates how precarious it is to infer pulse durations indirectly from partial knowledge of some of the parameters involved. Whether this deviation from the transform-limited value is due to temporal or spatial harmonic phase variation cannot be inferred from these measurements.

The present method can be extended to perform detailed parametric studies under various conditions, and refined to include a second-order frequency-resolved XUV gating by using energy resolved photoelectron detection²⁷, a full temporal characterization method analogous to frequency resolved optical gating (FROG). These improvements, complementary to the cross-correlation SPIDER^{28,29} or FROG type³⁰ approaches, will constitute a decisive step in the quest for ever-shorter, well-characterized and intense light pulses with a potential for advances in the field of attoscience. □

Received 23 May; accepted 29 September 2003; doi:10.1038/nature02091.

- Hänsch, T. W. A proposed sub-femtosecond pulse synthesizer using separate phase-locked laser oscillators. *Opt. Commun.* **80**, 71–75 (1990).
- Farkas, Gy & Tóth, Cs Proposal for attosecond light pulse generation using laser induced multiple-harmonic conversion processes in rare gases. *Phys. Lett. A* **168**, 447–450 (1992).
- Papadogiannis, N. A., Witzel, B., Kalpouzos, C. & Charalambidis, D. Observation of attosecond light localization in higher order harmonic generation. *Phys. Rev. Lett.* **83**, 4289–4292 (1999).
- Paul, P. M. *et al.* Observation of a train of attosecond pulses from high harmonic generation. *Science* **292**, 1689–1692 (2001).
- Christov, I. P., Murnane, M. M. & Kapteyn, H. High-harmonic generation of attosecond pulses in the 'single-cycle' regime. *Phys. Rev. Lett.* **78**, 1251–1254 (1997).
- Hentschel, M. *et al.* Attosecond metrology. *Nature* **414**, 509–513 (2001).
- Kobayashi, Y., Sekikawa, T., Nabekawa, Y. & Watanabe, S. 27-fs extreme ultraviolet pulse generation by high-order harmonics. *Opt. Lett.* **23**, 64–66 (1998).
- Papadogiannis, N. A. *et al.* Two XUV-photon ionization of He through a superposition of higher harmonics. *Phys. Rev. Lett.* **90**, 133902 (2003).
- Papadogiannis, N. A. *et al.* On the feasibility of performing non-linear autocorrelation with attosecond pulse trains. *Appl. Phys. B* **76**, 721–727 (2003).
- Kienberger, R. *et al.* Steering attosecond electron wave packets with light. *Science* **297**, 1144–1148 (2002).
- Drescher, M. *et al.* Time-resolved atomic inner-shell spectroscopy. *Nature* **419**, 803–807 (2002).
- Steinmeyer, G., Sutter, D., Gallmann, L., Matuschek, N. & Keller, U. Frontiers in ultrashort pulse generation: pushing the limits in linear and nonlinear optics. *Science* **286**, 1507–1512 (1999).
- Corkum, P. B. Plasma perspective on strong-field multiphoton ionization. *Phys. Rev. Lett.* **71**, 1994–1997 (1993).
- Lewenstein, M., Balcou, Ph., Ivanov, M. Yu., L'Huillier, A. & Corkum, P. B. Theory of high-harmonic generation by low-frequency laser fields. *Phys. Rev. A* **49**, 2117–2132 (1994).
- Salières, P. *et al.* Feynman's path-integral approach for intense-laser-atom interactions. *Science* **292**, 902–905 (2001).
- Chang, Z. *et al.* Temporal phase control of soft-x-ray harmonic emission. *Phys. Rev. A* **58**, R30–R33 (1998).
- Gaarde, M. B. *et al.* Spatiotemporal separation of high harmonic radiation into two quantum path components. *Phys. Rev. A* **59**, 1367–1373 (1999).
- Shin, H. J. *et al.* Nonadiabatic blueshift of high-order harmonics from Ar and Ne atoms in an intense femtosecond laser field. *Phys. Rev. A* **63**, 053407 (2001).
- Bartels, R. *et al.* Shaped-pulse optimization of coherent emission of high-harmonic soft X-rays. *Nature* **406**, 164–166 (2000).
- Christov, I. P., Bartels, R., Kapteyn, H. C. & Murnane, M. M. Attosecond time-scale intra-atomic phase matching of high harmonic generation. *Phys. Rev. Lett.* **86**, 5458–5461 (2001).
- Papadogiannis, N. A. *et al.* Temporal characterization of short pulse third-harmonic generation in an atomic gas by a transmission grating Michelson interferometer. *Opt. Lett.* **27**, 1561–1563 (2002).
- Wabnitz, H. *et al.* Multiple ionization of atom clusters by intense soft X-rays from a free-electron laser. *Nature* **420**, 482–485 (2002).
- Peatross, J., Chaloupka, J. L. & Meyerhofer, D. D. High-order harmonic-generation with an annular laser beam. *Opt. Lett.* **19**, 942–944 (1994).

24. Gaarde, M. B. & Schafer, K. J. Space-time considerations in the phase locking of high harmonics. *Phys. Rev. Lett.* **89**, 213901 (2002).
25. Constant, E., Mével, E., Zaïr, A., Bagnoud, V. & Salin, F. Toward sub-femtosecond pump-probe experiments: A dispersionless autocorrelator with attosecond resolution. *J. Phys. IV Fr.* **11**, Pr2-537–Pr2-540 (2001).
26. Mashiko, H., Suda, A. & Midorikawa, K. All-reflective interferometric autocorrelator for the measurement of ultra-short optical pulses. *Appl. Phys. B* **76**, 525–530 (2003).
27. Sekikawa, T., Katsura, T., Miura, S. & Watanabe, S. Measurement of the intensity-dependent atomic dipole phase of a high harmonic by frequency-resolved optical gating. *Phys. Rev. Lett.* **88**, 193902 (2002).
28. Muller, H. G. Reconstruction of attosecond harmonic beating by interference of two-photon transitions. *Appl. Phys. B* **74** (suppl.), S17–S21 (2002).
29. Quéré, F., Itatani, J., Yudin, G. L. & Corkum, P. B. Attosecond spectral shearing interferometry. *Phys. Rev. Lett.* **90**, 073902 (2003).
30. Norin, J. *et al.* Time-frequency characterization of femtosecond extreme ultraviolet pulses. *Phys. Rev. Lett.* **88**, 193901 (2002).

Acknowledgements The experiment was performed using the ATLAS laser facility at the Max-Planck-Institut für Quantenoptik, Garching. This work is supported in part by the European Community's Human Potential Programmes, Generation and Characterisation of Attosecond Pulses in Strong Laser-Atom Interactions: A Step towards Attophysics (ATTO) and the Ultraviolet Laser Facility (ULF).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.D.T. (george.tsakiris@mpq.mpg.de).

Drying-mediated self-assembly of nanoparticles

Eran Rabani¹, David R. Reichman², Phillip L. Geissler^{3*} & Louis E. Brus⁴

¹School of Chemistry, Tel Aviv University, Tel Aviv 69978, Israel

²Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA

³Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

⁴Department of Chemistry, Columbia University, 3000 Broadway, New York, New York 10027, USA

* Present address: Department of Chemistry, University of California, Berkeley, California 94720, USA

Systems far from equilibrium can exhibit complex transitory structures, even when equilibrium fluctuations are mundane^{1,2}. A dramatic example of this phenomenon has recently been demonstrated for thin-film solutions of passivated nanocrystals during the irreversible evaporation of the solvent^{3–14}. The relatively weak attractions between nanocrystals, which are efficiently screened in solution, become manifest as the solvent evaporates, initiating assembly of intricate, slowly evolving structures⁴. Although certain aspects of this aggregation process can be explained using thermodynamic arguments alone⁶, it is in principle a non-equilibrium process⁷. A representation of this process as arising from the phase separation between a dense nanocrystal 'liquid' and dilute nanocrystal 'vapour' captures some of the behaviour observed in experiments³, but neglects entirely the role of solvent fluctuations, which can be considerable on the nanometre length scale¹⁵. Here we present a coarse-grained model of nanoparticle self-assembly that explicitly includes the dynamics of the evaporating solvent. Simulations using this model not only account for all observed spatial and temporal patterns, but also predict network structures that have yet to be explored. Two distinct mechanisms of ordering emerge, corresponding to the homogeneous and heterogeneous limits of evaporation dynamics. Our calculations show how different choices of solvent, nanoparticle size (and identity) and thermodynamic state give rise to the various morphologies of the final

structures. The resulting guide for designing statistically patterned arrays of nanoparticles suggests the possibility of fabricating spontaneously organized nanoscale devices.

In our model, the solvent is represented as a two-dimensional lattice gas¹⁶. In detail, each cell of a square lattice, with the size of the solvent's correlation length, $\xi \approx 1$ nm, is occupied by either liquid ($l_i = 1$) or vapour ($l_i = 0$), where l_i is a binary variable roughly proportional to solvent density at lattice site i . Adjacent liquid cells attract one another with a strength ϵ_l determined by the energy density of the fluid. The chemical potential, μ , and temperature, T , of a surrounding bath (in experiments, the overlying vapour) establish the average concentration of liquid and vapour cells at equilibrium. We consider non-equilibrium conditions appropriate for experiments conducted in a closed chamber, in which small amounts of solvent may remain on the surface long after the majority has evaporated. Initially, the lattice is entirely filled with liquid and nanoparticles. But μ is fixed at a value for which the equilibrium state is vapour. Although the subsequent decay of solvent density is neither constant nor simply exponential for this choice of ensemble, we find that it is well characterized by a single timescale. (See Supplementary Information.)

A cell of the lattice that is not occupied by liquid may instead be occupied by a nanoparticle. This fact, together with attractions between neighbouring liquid and nanoparticles (of strength ϵ_{nl}), can alter the drying dynamics significantly. We use a second set of binary variables, n_i , to describe nanoparticle density on the lattice. Where nanoparticles exclude solvent, $n_i = 1$, and elsewhere, $n_i = 0$. Because the size of a nanoparticle, d , can exceed the range of correlated solvent fluctuations, we allow them to span several cells of the lattice. Adjacent nanoparticle cells attract one another, with strength ϵ_n , so that they tend to aggregate in the absence of liquid. The geometry and relevant length scales of this lattice model are sketched in Fig. 1.

The dynamics of our model are stochastic, both for fluctuations in solvent density and for nanoparticle diffusion. In the former case, configurations evolve by Monte Carlo dynamics. We attempt to convert a randomly chosen lattice cell, i , from liquid to vapour (or vice versa), $l_i \rightarrow (1 - l_i)$. This perturbation is accepted with a

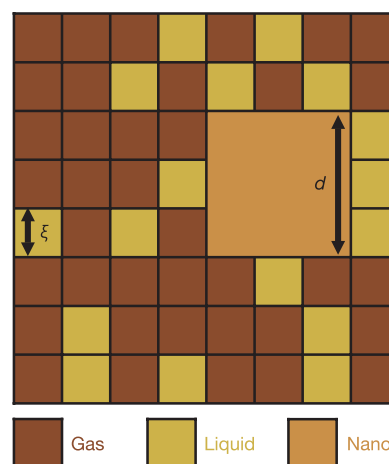


Figure 1 A sketch of the square lattice and important length scales of our mesoscopic model. Each lattice cell of size ξ is occupied by gas, liquid or nanoparticle. The colour scheme shown is maintained in subsequent figures. For results in those figures, we have taken $\epsilon_n = 2\epsilon_l$ and $\epsilon_{nl} = 1.5\epsilon_l$, for which nanoparticles are well solvated before evaporation. The lattice comprises $1,000 \times 1,000$ cells unless otherwise stated, and each nanoparticle spans 4×4 cells. (As observed in experiments^{3,4}, the qualitative results of our model are not sensitive to the size of the nanoparticles.)

Metropolis probability, $p_{\text{acc}} = \min [1, \exp (-\Delta H/k_B T)]$, where k_B is Boltzmann's constant, and ΔH is the resulting change in energy. For a particular configuration of $\{l_i\}$ and $\{n_i\}$:

$$H = -\epsilon_l \sum_{\langle ij \rangle} l_i l_j - \epsilon_n \sum_{\langle ij \rangle} n_i n_j - \epsilon_{nl} \sum_{\langle ij \rangle} n_i l_j - \mu \sum_i l_i \quad (1)$$

The first three summations in equation (1) include only adjacent lattice cells. These dynamics do not conserve solvent density within the film, as seems appropriate for the experiments under study. Nanocrystals, on the other hand, have negligible vapour pressure, and their surface density should be locally conserved. In our model, nanoparticles execute a random walk on the lattice, biased by their interactions with liquid cells and with each other. We attempt to displace a nanoparticle by a single lattice spacing in a randomly chosen direction. Such a move is accepted also with the above Metropolis probability, but only if the region into which the nanoparticle moves is completely filled with liquid. Solvent density in lattice cells overtaken by this displacement is regenerated in the wake of the moving nanoparticle. This final constraint mimics the very low mobility of nanocrystals on a dry surface. It also provides an additional coupling between the kinetics of evaporation and nanoparticle phase separation.

In the absence of solvent fluctuations, our model of nanoparticle self-assembly is related to simple models of coarsening that have been studied extensively in the contexts of ferromagnetic phase-ordering and spinodal decomposition^{1,2}. When, for example, an Ising magnet is quenched below its critical temperature, amorphous but nearly homogeneous domains form rapidly and then coalesce. This union of ever-larger domains proceeds in essentially the same way at each stage, but with a growing characteristic length scale, $R(t)$. The coarsening is thus self-similar, and $R(t)$ evolves as $R \approx t^\alpha$. Exponents for this dynamical power law depend on the nature of dynamics and physical constraints¹.

Our model should reflect this basic phenomenology in certain limits of the parameters controlling solvent evaporation and nanoparticle aggregation. The appropriate limit is not one of instantaneous evaporation, because nanoparticles are immobile in the

absence of solvent. We find that simple coarsening behaviour is instead realized when evaporation is spatially homogeneous while the boundaries of nanoparticle domains remain fluxional throughout the growth dynamics. The latter condition requires that boundaries are wet (that is, covered with solvent) and that the energetic cost of moving a nanoparticle into the surrounding solvent is comparable to $k_B T$. In this limit the scarcity of solvent drives aggregation of nanoparticles uniformly, while growing domains are free to rearrange and conglomerate. We expect this scenario when the solvent is near the spinodal limit of metastability, $\mu \approx \mu_{\text{sp}}(T)$, while a wetting layer remains thermally stable at the edges of nanoparticle domains, $\mu + \epsilon_{nl} + \epsilon_l > k_B T$.

A survey of aggregate morphologies obtained in this limit of homogeneous drying is provided in Fig. 2. The lower panels show snapshots from simulated non-equilibrium trajectories, at intermediate times. The chosen trajectories differ from one another only in coverage (that is, mean surface density) and mobility of nanoparticles. At low coverage (Fig. 2a), distinct, disk-like aggregates of

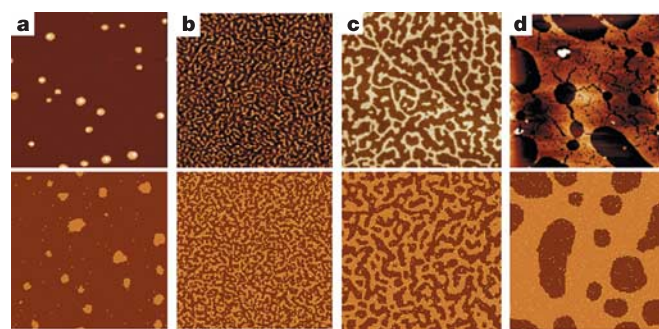


Figure 2 Self-assembled morphologies resulting from homogeneous evaporation and wetting of nanoparticle domains. The lower panels show results of model simulations for coverages of 5% (a), 30% (b), 40% (c) and 60% (d). In all cases $k_B T = \epsilon_l/2$ and $\mu = -2.25\epsilon_l$. ($\mu = -2\epsilon_l$ corresponds to phase coexistence of liquid and vapour.) Although the characteristic times of evaporation, τ , and nanoparticle diffusion, τ_D , are increasing functions of coverage, their ratio, $\tau/\tau_D \approx 200$, is nearly constant. The size of the domains is largely determined by the nanoparticle mobility and the time window, τ_s , before growth is stopped. Disks containing several hundred nanoparticles form for the lowest coverage at $\tau_s/\tau \approx 2,500$. In b, small domains form at $\tau_s/\tau \approx 4$. Larger aggregates appear by $\tau_s/\tau \approx 50$ and $\tau_s/\tau \approx 500$ in c and d, respectively. The upper panels show corresponding experimentally observed morphologies for CdSe nanocrystals at similar coverages. Solvents used in these experiments range from hexane (a and c) to chloroform (b) and a 90:10 mixture of chloroform and methanol (d). Comparison with simulation results is therefore strictly qualitative.

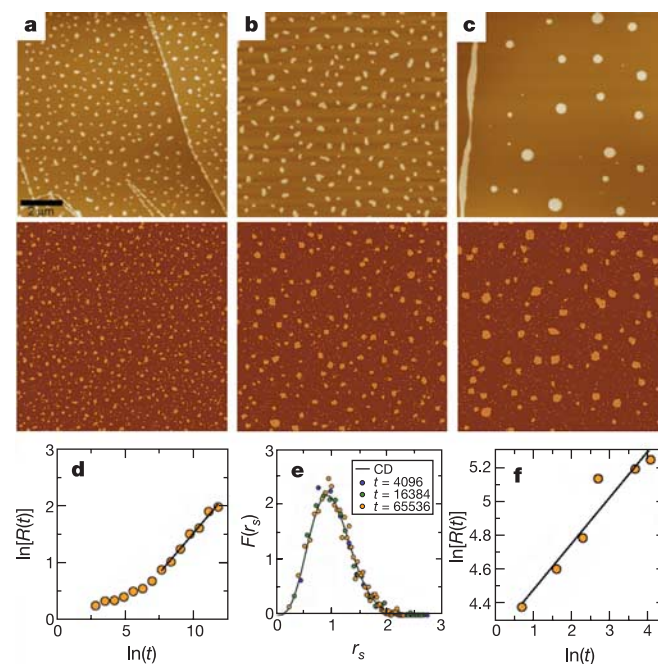


Figure 3 Dynamics of nanoparticle assembly at low coverage. a–c, The lower panels show simulation results for the same thermodynamic state as in Fig. 2, with a time range of $\tau_s = 15\tau$ (a) to $\tau_s = 600\tau$ (c). The upper panels depict experimental results for PbSe nanocrystals in octane, with a range of $\tau_s = 2$ min (a) to $\tau_s = 60$ min (c) in the time elapsed before sudden exposure to the atmosphere induced complete drying. In calculations, evaporation is slow compared to nanoparticle mobility, $\tau/\tau_D \approx 1,000$. We assign a physical timescale, Δt , to a single Monte Carlo step in our simulations by identifying the mean squared displacement of a nanoparticle (under non-evaporating conditions) after m steps as $4D_{\text{exp}}m\Delta t$, where D_{exp} is a diffusion coefficient estimated from experimental data. With a physical length scale of $\xi = 1$ nm (yielding an overall scale of $1 \mu\text{m}$ for the depicted simulation boxes), the time dependence of domain growth in our model closely mirrors that of experiments. The very long timescales corresponding to the upper panels of (a–c) are not accessible by our simulations. A side-by-side comparison is made possible by the self-similar nature of these dynamics. The logarithm of the average domain size, $R(t)$, is plotted as a function of $\ln t$ for theory (d) and experiment (f). Solid lines, with slope 0.28 (d) and 0.27 (f), are best linear fits to the last several data points. Distributions of scaled disk sizes, $r_s = r/R(t)$, where r is the size of a particular disk, are plotted in e for several times, t , in the simulations depicted in a–c. Each distribution is averaged over several non-equilibrium trajectories. Assuming that disk mobility is proportional to $1/r$, cluster diffusion (CD) theory predicts the time-independent distribution shown in e as a solid line.

nanoparticles dominate the self-assembly. At higher coverage (Fig. 2b and c), nanoparticle domains are anisotropic and nearly percolate through the lattice. For very long times, however, even these fingering, high-coverage patterns tend toward large disks (Fig. 2d) and, asymptotically, to a single, macroscopic disk (not shown in the figure). Disk-like and ribbon-like morphologies thus represent different stages of the same growth mechanism.

The upper panels of Fig. 2 are micrographs of nanoparticle self-assembly in our experiments. In each case, the nanoparticle density lies within two per cent of the simulated system depicted in the adjacent panel, so that the dependence of pattern formation on coverage can be compared. With a reasonable estimate of the physical timescale for computed trajectories, the growth behaviour in simulation and experiment agree well. Comparing the detailed conditions of theory and experiment is made difficult by the lack of thermodynamic data for relevant thin-film solvents. We have been careful, however, to vary only those parameters that are changed or uncontrolled in the laboratory. Specifically, temperature, nano-

particle size and attraction are held fixed. The use of different solvents and evaporation conditions effectively adjusts chemical potential, nanoparticle mobility, and the attractions ϵ_1 and ϵ_{nl} . We have chosen these model parameters within physically reasonable limits.

The strong visual correspondence between the experimental and theoretical results in Fig. 2 demonstrates that our coarse-grained model successfully captures the qualitative dynamics of nanoparticle assembly in the limit of spatially homogeneous evaporation. A more quantitative comparison is illustrated in Fig. 3, depicting the time evolution of nanoparticle aggregation. Not only are coalescence and shape fluctuations visually similar, but the scaling of domain size with time is essentially identical. Following a transient period, $R(t)$ grows as t^α with $\alpha \approx 1/4$. This exponent is consistent with growth by cluster diffusion (in which domains grow solely by coalescence), assuming that cluster mobility scales as R^{-1} . Figure 3e shows a more sensitive signature of this mechanism, namely, scaling behaviour of the cluster size distribution. These dynamics have been shown to dominate growth at intermediate times for a model similar to ours, but without solvent fluctuations or kinetic constraints¹⁸. We thus conclude that when evaporation is spatially uniform, the coarsening of nanoparticle domains is essentially that of an appropriate one-component fluid, as suggested in refs 3 and 4. Solvent evaporation simply triggers the onset of coarsening, after which the mechanism and timescale of growth are determined by nanoparticle diffusion.

When evaporation is instead strongly heterogeneous in space, the dynamics of self-assembly can be dramatically different. Between binodal and spinodal lines, $-2\epsilon_1 > \mu > \mu_{sp}(T)$, solvent remains locally metastable on the surface. In this case phase change occurs not by long-wavelength fluctuations, but by nucleation and growth of vapour bubbles. At a given instant, the driving force for nanoparticle assembly therefore varies with position on the lattice. If nanoparticles are sufficiently mobile to track the fronts of growing vapour nuclei, their aggregate patterns will be shaped by the structural history of evaporation. Specifically, the locations of nanoparticle domains at long times should roughly trace the intersection lines of colliding vapour nuclei, leading to network-like morphologies. This assembly mechanism is similar in many respects to the Marangoni effect¹⁹. But in our model, the effective pinning of solutes to vapour fronts is not hydrodynamic in nature. Rather, it is a direct consequence of the coupling between two phase transitions, one in solvent density and the other in nanoparticle density.

Figure 4 shows a few representative simulation trajectories and experimental morphologies under conditions of heterogeneous evaporation. Figure 4a–c exemplifies network formation when domain edges are effectively frozen after evaporation, so that aggregation essentially halts when vapour nuclei meet. Each cell of the network structures in Fig. 4b and c thus marks an independent nucleation event whose front pushed nanoparticles to the boundaries of adjacent cells. The shapes of terminal structures in these trajectories are determined primarily by the relative timescales of evaporation, τ , and nanoparticle motion, $\tau_D = \xi^2/D$, where D is the nanoparticle diffusion coefficient in solution. For large τ/τ_D , the edges of nanoparticle domains may rearrange before the disappearance of solvent renders them immobile. The smooth edges of rod-like aggregates in Fig. 4a and of network cell boundaries in Fig. 4b are the result of such relaxation. By contrast, when τ/τ_D is small, domain edges are effectively frozen as they are formed, resulting in highly ramified networks such as that in Fig. 4c. The fractal-like appearance of these structures originates in dynamics that are locally similar to diffusion limited aggregation²⁰. Attachment of nanoparticles to growing domains is nearly irreversible as the front of a vapour nucleus passes by. Strictly irreversible binding of random walkers is known to generate fractal structures. Cellular morphologies have been observed in several experiments^{3,19,21}. One

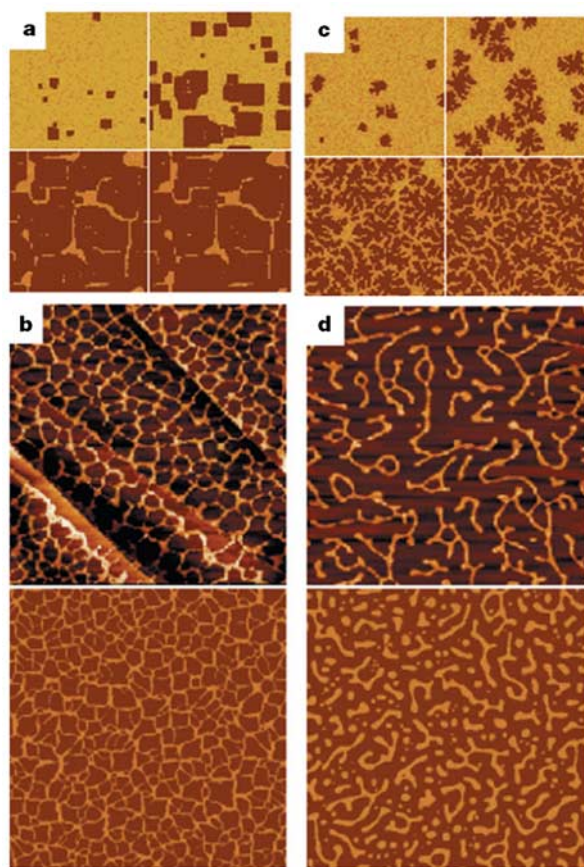


Figure 4 Self-assembled morphologies resulting from inhomogeneous evaporation in simulations and in experiments. Simulations were performed at the same temperature and chemical potential as in Fig. 2, but with stronger attractions between solvent cells, $\epsilon_1 = 4k_B T$. Results are shown for coverages of 10% (**a**), 20% (**b**, bottom, and **d**, bottom), and 30% (**c**). In panels **a–c** the domain edges are not fluxional. Evaporation times are comparable in these cases, but nanoparticle mobility is varied significantly, yielding $\tau/\tau_D \approx 6,500$ in **a**, $\tau/\tau_D \approx 800$ in **b**, bottom, and $\tau/\tau_D \approx 150$ in **c**. A network structure observed in experiments with CdSe nanocrystals in hexane is shown **b**, top. The corresponding simulation result is shown in **b**, bottom. When domain edges remain fluxional following heterogeneous evaporation, cells break up as diffusion concentrates nanoparticle density at the nodes of the network, leaving distinct, worm-like domains as shown in **d**, bottom. In **d**, top, we show worm-like structures observed in experiments with CdSe nanocrystals in chloroform. The simulation results in **b** and **d** were obtained with a larger lattice ($4,000 \times 4,000$).

such structure is depicted in the upper panel of Fig. 4b. It compares very well qualitatively with the model result shown in the lower panel of Fig. 4b.

When domain edges remain fluxional following heterogeneous evaporation, the networks described above for Fig. 4b are not stable, long-lived structures. Nanoparticles continue to move in this case, strongly biased by the interfacial tension of cell boundaries. Cells break up as diffusion concentrates nanoparticle density at the nodes of the network, leaving distinct, worm-like domains. An example of such a pattern generated by our simulations (lower panel of Fig. 4d) compares well with worm-like morphologies observed in experiments (upper panel of Fig. 4d). These structures are themselves transitory, because their anisotropy costs significant interfacial free energy. Domains thus eventually become disks, which diffuse and coalesce, as described in the case of homogeneous evaporation. This mechanism of network disintegration strongly resembles that observed in viscoelastic phase separation of a dynamically asymmetric mixture².

Our results suggest four basic regimes of drying-mediated nanoparticle assembly. They are distinguished by the spatial uniformity of solvent dynamics, and by the fluctuations of nanoparticle domain boundaries following evaporation. When solvent disappears homogeneously from the surface, disk-like or ribbon-like domains reminiscent of spinodal decomposition form at early times. We have shown that if these aggregates remain fluxional, they continue to evolve in a self-similar fashion, principally by diffusion and coalescence. If instead domain boundaries are frozen following evaporation, dynamical constraints arrest this growth at an early stage. When evaporation is inhomogeneous owing to infrequent nucleation events, network structures are formed at early times as vapour nuclei meet. These cellular patterns are only stable if interfaces are frozen following evaporation. Otherwise, networks fragment to form distinct domains that asymptotically evolve as in homogeneous coarsening.

Many aspects of these self-assembly mechanisms have previously been rationalized on the basis of seemingly distinct physical pictures. Our mesoscopic model unifies these pictures and provides a quantitative measure of their importance under different conditions. As such, it may serve as a basic guide for designing self-assembled structures with desired nanometre-scale features. Considering the simplicity of its energetics and dynamical rules, the degree of local order and anisotropy out of equilibrium is remarkable. We have omitted many physical details, most notably hydrodynamic convection, substrate roughness, non-local interactions, and film thickness. These effects can be added to provide microscopic realism, but it is interesting that they are not needed to account for the variety of patterns that have been observed in experiments. □

Received 9 May; accepted 25 September 2003; doi:10.1038/nature02087.

- Bray, A. J. Theory of phase-ordering kinetics. *Adv. Phys.* **43**, 357–459 (1994).
- Tanaka, H. Viscoelastic phase separation. *J. Phys. Condens. Matter* **12**, R207–R264 (2000).
- Ge, G. & Brus, L. E. Evidence for spinodal phase in two-dimensional nanocrystal self-assembly. *J. Phys. Chem. B* **104**, 9573–9575 (2000).
- Tang, J., Ge, G. & Brus, L. E. Gas-liquid-solid phase transition model for two-dimensional nanocrystal self-assembly on graphite. *J. Phys. Chem. B* **106**, 5653–5658 (2002).
- Puntes, V. F., Krishnan, K. M. & Alivisatos, A. P. Colloidal nanocrystal shape and size control: The case of cobalt. *Science* **291**, 2115–2117 (2001).
- Gelbart, W. M., Sear, R. P., Heath, J. R. & Chaney, S. Array formation in nano-colloids: Theory and experiment in 2D. *Farad. Disc.* **112**, 299–307 (1999).
- Whitesides, G. M. & Grzybowski, B. Self-assembly at all scales. *Science* **295**, 2418–2421 (2002).
- Murray, C. B., Kagan, C. R. & Bawendi, M. G. Self-organization of CdSe nanocrystallites into 3-dimensional quantum-dot superlattices. *Science* **270**, 1335–1338 (1995).
- Freeman, R. G. *et al.* Self-assembled metal colloid monolayers—an approach to SERS substrates. *Science* **267**, 1629–1632 (1995).
- Andres, R. P. *et al.* Self-assembly of a two-dimensional superlattice of molecularly linked metal clusters. *Science* **273**, 1690–1693 (1996).
- Harfenist, S. A., Wang, Z. L., Alvarez, M. M., Vezmar, I. & Whetten, R. L. Highly oriented molecular Ag nanocrystal arrays. *J. Phys. Chem.* **100**, 13904–13910 (1996).
- Sear, R. P., Chung, S. W., Markovich, G., Gelbart, W. M. & Heath, J. R. Spontaneous patterning of quantum dots at the air-water interface. *Phys. Rev. E* **59**, R6255–R6258 (1999).

- Fried, T., Shemer, G. & Markovich, G. Ordered two-dimensional arrays of ferrite nanoparticles. *Adv. Mater.* **13**, 1158–1161 (2001).
- Redl, F. X., Cho, K. S., Murray, C. B. & O'Brien, S. Three-dimensional binary superlattices of magnetic nanocrystals and semiconductor quantum dots. *Nature* **423**, 968–971 (2003).
- Elbaum, M. & Lipson, S. G. How does a thin wetted film dry up? *Phys. Rev. Lett.* **72**, 3562–3565 (1994).
- Chandler, D. *Introduction to Modern Statistical Mechanics* (Oxford Univ. Press, New York, 1987).
- Ge, G. & Brus, L. E. Fast surface diffusion of large disk-shaped nanocrystal aggregates. *Nano Lett.* **1**, 219–222 (2001).
- Lo, A. & Skoedje, R. T. Kinetic and Monte Carlo models of thin film coarsening: Cross over from diffusion-coalescence to Ostwald growth modes. *J. Chem. Phys.* **112**, 1966–1974 (2000).
- Maillard, M., Motte, L., Ngo, A. T. & Pileni, M. P. Rings and hexagons made of nanocrystals: A Marangoni effect. *J. Phys. Chem. B* **104**, 11871–11877 (2000).
- Witten, T. A. & Sander, L. M. Diffusion-limited aggregation, a kinetic critical phenomenon. *Phys. Rev. Lett.* **47**, 1400–1403 (1981).
- Stowell, C. & Korgel, B. A. Self-assembled honeycomb networks of gold nanocrystals. *Nano Lett.* **1**, 595–600 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by the United States–Israel Binational Science Foundation. L.E.B. is supported by the Columbia MRSEC. P.L.G. was an MIT Science Fellow throughout most of this work. D.R.R. is a Sloan Fellow and Camille Dreyfus Teacher-Scholar.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.R. (rabani@tau.ac.il) or D.R.R. (reichman@chemistry.harvard.edu).

Proxy evidence for an El Niño-like response to volcanic forcing

J. Brad Adams¹, Michael E. Mann¹ & Caspar M. Ammann²

¹Department of Environmental Sciences, University of Virginia, Clark Hall, Charlottesville, Virginia 22903, USA

²Climate Global Dynamics Division, National Center for Atmospheric Research, 1850 Table Mesa Drive, Boulder, Colorado 80307-3000, USA

Past studies have suggested a statistical connection between explosive volcanic eruptions and subsequent El Niño climate events^{1,2}. This connection, however, has remained controversial^{3–5}. Here we present support for a response of the El Niño/Southern Oscillation (ENSO) phenomenon^{6,7} to forcing from explosive volcanism by using two different palaeoclimate reconstructions of El Niño activity^{8,9} and two independent, proxy-based chronologies of explosive volcanic activity⁵ from AD 1649 to the present. We demonstrate a significant, multi-year, El Niño-like response to explosive tropical volcanic forcing over the past several centuries. The results imply roughly a doubling of the probability of an El Niño event occurring in the winter following a volcanic eruption. Our empirical findings shed light on how the tropical Pacific ocean–atmosphere system may respond to exogenous (both natural and anthropogenic) radiative forcing.

Coupled ocean–atmosphere experiments have explored the possible response of ENSO to enhanced greenhouse gas concentrations^{10–16}. Results indicate El Niño-like^{11–14}, neutral¹⁵ and even La Niña-like¹⁶ responses of average conditions (even a ‘neutral’ response represents a La Niña-like anomaly in the face of large-scale greenhouse warming). Simulations employing the Cane–Zebiak model of tropical Pacific coupled ocean–atmosphere dynamics, which exhibits a stronger dynamical feedback than most global models, produces negative (positive) eastern tropical Pacific sea surface temperature (SST) anomalies in response to a positive (negative) surface radiative forcing¹⁷. This imposes a La Niña-like cooling of the mean state in the presence of positive greenhouse warming¹⁰. It is intriguing in this context to reconsider the con-

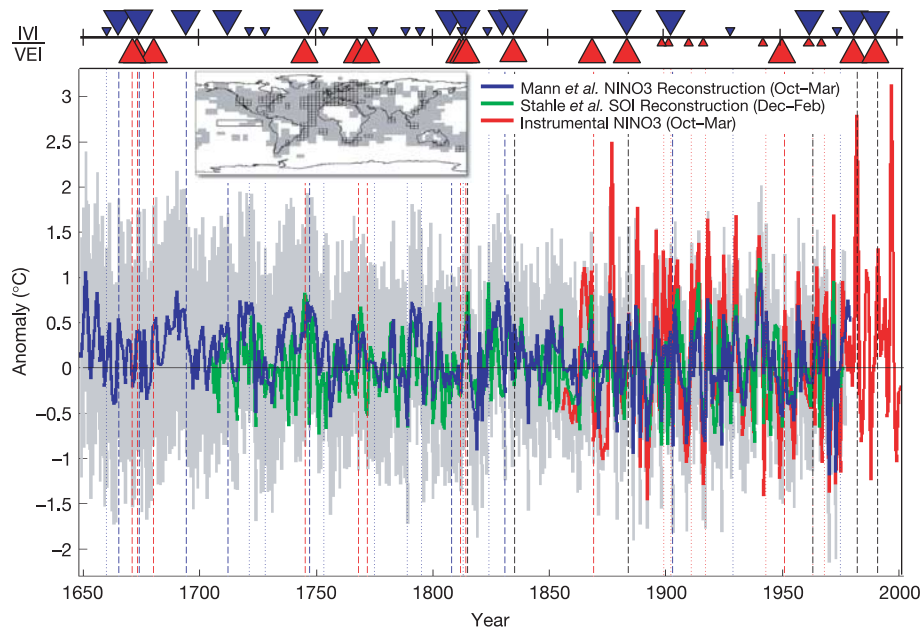


Figure 1 Long-term behaviour of ENSO. The Mann *et al.*⁹ cold-season (October–March) NINO3 (relative to 1901–80) (solid blue line) and Stahle *et al.*⁸ winter (December–February) SOI (solid green line) reconstruction are shown. The SOI is inverted to exhibit a positive correlation with ENSO events, and scaled to the standard deviation of the Mann *et al.* NINO3 reconstruction. The instrumental cold-season NINO3²⁸ is shown for comparison (solid red). Large (small) triangles and dashed (dotted) vertical lines denote

cold seasons (defined by beginning year of the cold season) following large (moderate) VEI or IVI eruptions back to 1649. VEI (IVI) eruptions marked by red (blue) arrows/vertical lines. Black vertical dashed lines show eruptions indicated by VEI and IVI. The inset shows reconstructed surface temperature grids in the NINO3 region (small rectangle) used to calculate the Mann *et al.* index. Grey shading indicates the 95% confidence limits of the NINO3 reconstruction⁹.

controversial findings of Handler and Andsager^{1–4} arguing for a relationship between explosive, tropical volcanic eruptions and El Niño events. Large eruptions can inject aerosols into the lower tropical stratosphere, imparting a significant dust veil with a persistent (typically one to three year) associated negative short-wave surface radiative forcing. This radiative forcing should, by the reasoning of refs 10 and 17, lead to a multi-year, El Niño-like response of the tropical Pacific ocean–atmosphere system. Statistical validation of the relationship outlined by Handler and colleagues may have broader implications for ENSO’s response to greenhouse warming¹⁰. In turn, the demonstrated plausibility of a mechanism for an ENSO-like response to radiative forcing¹⁷ combined with our analysis addresses a major objection to the Handler findings: although previously suggested physical mechanisms¹⁸ could not be confirmed¹⁹, only now does a compelling mechanism have theoretical and empirical support.

Additional criticisms of past work^{1,2} include the subjectivity of the applied volcanic chronology^{5,20}, the small number (12) of eruptions used (thereby limiting statistical robustness), and other methodological deficiencies^{21–24}. In particular, the existence of a causally consistent relationship has been called into question as El Niño events precede the actual eruptions for some events^{3,4}. Indeed, the instrumental period probably provides unsatisfactory statistical constraint for determining if a true relationship exists between explosive volcanism and El Niño variability. To reassess the ‘volcano–El Niño’ hypothesis independently of past work, we rely on much longer proxy-based reconstructions of both forcing (derived histories of explosive tropical eruptions) and response (proxy-based diagnostics of ENSO). In the process, we also address several shortcomings of the original statistical methodology^{21–24}.

We make use of two independent measures of past volcanic activity—the volcanic explosivity index (VEI), and (2) the ice-core volcanic index (IVI). The VEI estimates volcanic explosivity for events in the comprehensive Smithsonian catalogue. It is based partially on qualitative volcanological information and should

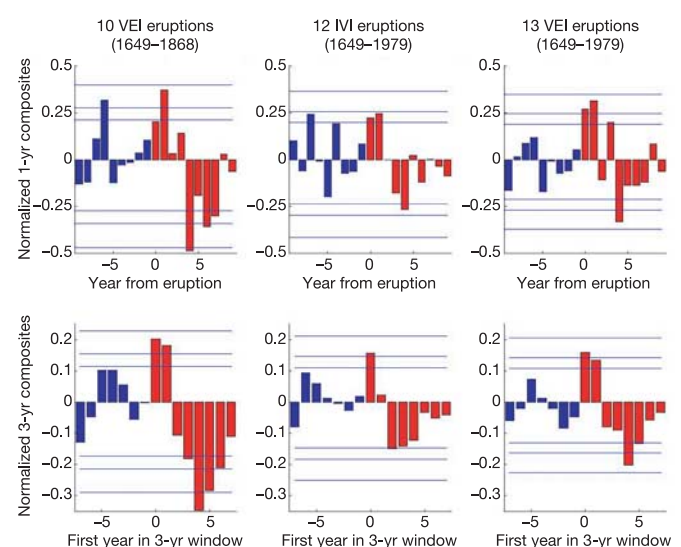


Figure 2 Reconstructed, cold-season NINO3 SEA results. Normalized results for ten pre-instrumental (1649–1868) VEI eruptions (leftmost panels) and the 12 (13) largest IVI (VEI) eruptions for the full period (1649–1868) (middle and rightmost panels—both include post-eruption cooling removal) are shown. Upper (lower) panels show 1-yr (3-yr) composites. Annual composites are used to generate 3-year moving averages (that is, the 3-up composites). Confidence limits (90%, 95%, 99%) are marked by horizontal lines. Red and blue colours mark the post-eruption and pre-eruption composites. Five of seven pre-instrumental analyses exhibit a post-eruption El Niño-like response. Similar results are obtained using the full period using IVI and VEI indices. Substituting the DVI as the volcanic index or the SOI as the ENSO diagnostic yields similar results. Instrumental analyses (1869–1979) including all Handler eruptions^{1,2} or a moderately selective list of (11) VEI-based instrumental eruptions also indicate a significant El Niño-like response. Analyses using Handler’s eruptions (except El Chichón and one other) do not, however, show an El Niño-like response (Supplementary Information). The instrumental period yields variable SEA results.

therefore be used with caution in studies seeking quantitative results^{5,25,26}. The IVI, in contrast, is tied more closely to the injection of climatically relevant volcanic sulphate aerosols into the lower stratosphere that can be recovered from polar ice^{20,26}. An eruption is assumed to have occurred in the tropics if its aerosols are recorded at both poles. The precise timing of each eruption is, however, somewhat uncertain because the aerosols' residence time in the atmosphere varies owing to temporally and spatially variable atmospheric factors and inherent dating problems within the cores themselves. We discretize the IVI into a '1' (smallest) to '8' (largest) scale parallel to the VEI. We consider an event magnitude of 4 to be the minimum magnitude of climatic relevance, and event magnitudes of 5 or greater are used to classify the largest of climatically relevant events. Another available volcanic index is the dust veil index (DVI), but because of the potential circularity that arises in using this index²⁰ we focus here on the VEI and IVI. Excellent critiques of these volcanic indices and their application to studies of climate forcing are available^{20,24}. Overall, because of general agreement between all three indices in regard to the very large eruptions, all results presented appear robust with respect to the particular eruption index employed.

We also make use of two conventional diagnostics of ENSO—the Southern Oscillation index (SOI) and the NINO3 index (see Fig. 1). Both are evaluated during the boreal winter (December–February) or cold (October–March) season during which ENSO events are most pronounced. The instrumental SOI index²⁷, available back to 1866, measures atmospheric circulation changes associated with ENSO through the standardized difference in atmospheric surface pressure at Tahiti in the South Pacific Ocean and Darwin, Australia. El Niño events are generally associated with negative SOI values. The instrumental NINO3 index²⁸, available back to 1856, measures relative temperature variations in the eastern tropical Pacific Ocean associated with ENSO. The instrumental winter SOI and cold-season NINO3 indices are highly anticorrelated ($r \approx -0.7$).

We reassess previous findings using a reconstruction of the cold-season NINO3 series based on proxy climate indicators (tree rings,

ice cores, corals and historical documents), which exhibits considerable skill in cross-validation with independent early instrumental data⁹, and is available back to AD 1649. We make use of an additional proxy-based reconstruction of the winter SOI index based on ENSO-sensitive Mexican and southwestern US tree-ring data⁸, available back to 1706. The two reconstructions, which are largely, though not entirely, independent (both use a small subset of common proxy data), are strongly anticorrelated ($r = -0.63$) at a level similar to that seen between the corresponding instrumental series. This suggests that the reconstructions represent reasonably faithful, if uncertain, estimates of the respective indices several centuries back. The volcanic forcing and ENSO response data are incorporated into a superposed epoch analysis (SEA; see Methods), which is used to evaluate the composite response of ENSO to volcanic forcing events.

We first produced mildly significant results similar to those of Handler and colleagues^{1,2} by using their key dates and instrumental cold-season NINO3 data. When excluding several eruptions (for example, El Chichón) from Handler's eruption list, however, their results were not reproduced (see Supplementary Tables S1–S3). This underscores the potential leverage of a few events in analyses of the instrumental record, and the consequent fundamental sample size limitations in reaching definitive conclusions from the instrumental record alone⁷. We then performed SEA experiments over the extended 'pre-instrumental' intervals, for a variety of selection criteria, using both the IVI and VEI eruption lists, and both the reconstructed SOI and NINO3 series. Several VEI- and IVI-based SEA results during the pre-instrumental and full periods are shown in Fig. 2. Detailed SEA results are provided in Supplementary Information (Tables S1–S5; note that not all analyses are statistically independent, because of overlap between some eruption lists). The analyses generally indicate a positive, El Niño-like NINO3 composite response in the several years following explosive low-latitude eruptions. In the 20 different NINO3 analyses, 17 out of 25 of the positive annual responses significant at the $P = 0.05$ level occurred within the first three years (0, +1, +2) and only 8 before the

Table 1 **Reconstructed NIN03**

Event criteria	Period	3-yr SEA results															Count fraction
		<i>n</i>	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	
IVI Medium/Large	1649–1979	25										N					9/25
IVI Medium/Large	1649–1868	18							P		N	N	N				7/18
IVI Medium/Large	1706–1868	13							P				N				5/13
IVI Medium/Large	1706–1977	19															7/19
IVI Largest	1649–1979	12							P								6/12
IVI Largest	1706–1977	8							P								4/8
VEI ≥ 4	1649–1979	20							P				N*	N*			11/20
VEI ≥ 4	1869–1979	9															5/9
VEI ≥ 4	1649–1868	10							P	P			N*	N			4/10
VEI ≥ 4	1706–1977	16							P	P			N	N*	N	N	8/16
VEI ≥ 4	1706–1868	7							P*	P*			N	N	N		2/7
VEI Medium/Large	1649–1979	31							P*								15/31
Handler List	1869–1979	10															2/10
VEI Medium/Large	1869–1979	9															5/9
VEI Medium/Large	1649–1868	22							P*								9/22
VEI Medium/Large	1706–1868	18	P			N			P	P							5/18
VEI Medium/Large	1706–1977	26							P*	P*			N	N		N	10/26
VEI Largest	1649–1979	13							P				N				6/13
VEI Largest	1706–1977	10							P								4/10
VEI Largest	1706–1868	7							P*	P*			N	N	N		2/7
Positive response	>95%		1	0	0	0	0	0	0	15	6	0	0	0	0	0	Avg: 6.3/15.2
Negative responses	<95%		0	0	0	1	0	0	0	0	0	1	2	9	6	3	2
Combined total at	95%				2									44			
Positive response	>99%		0	0	0	0	0	0	0	5	3	0	0	0	0	0	
Negative responses	<99%		0	0	0	0	0	0	0	0	0	0	0	2	2	0	0
Combined total at	99%				0									12			

Table shows normalized, mean-adjusted 3-yr SEA and 'count' results. Years with positive (negative) 3-yr SEA composites at the 95 and 99% confidence level are marked by P and P* (N and N*), respectively. Each 3-yr composite is labelled with the first year in the 3-yr window moving away from the eruption year (for example, year -1 = mean of years -1, -2, -3; year 0 = mean of years 0, 1, 2). The last two columns display the fraction of cold-season anomalies in year +1 exhibiting an El Niño event (defined as an anomaly $>+0.3^{\circ}\text{C}$ relative to the local mean) for each key list. The mean count result (6.3/15.2) corresponds to a ~42% occurrence of El Niño's following eruptions. Comparatively, the expected mean percentage of El Niño events based on application of this definition to the long-term record is 24.8%. Significance determination is described in Methods. Count fraction significance values are shown in boldface: $P < 0.05$.

eruptions. Post-event El Niño-like warming is even more consistent if the multi-year (3-yr) composites are considered. With 21 of 22 positive responses significant at the $P = 0.05$ level (and all 8 of the $P = 0.01$ significant positive responses) found in the three-year windows starting in years 0 and 1, we infer that a preferred multi-year El Niño-like response to tropical volcanic forcing events is most consistent with the results.

Interestingly, a reversal of the SST response to statistically significant negative (La Niña-like) values following the initially observed El Niño-like response is also clearly observed (Fig. 2). The initial (El Niño) and secondary (La Niña) post-eruption composite responses (that is, year 0 to year 7) contain 42 of 44 three-year composites significant at the 95% confidence level, and all 12 at 99%. Of 24 negative responses significant at 95% in the 3-yr analysis, 18 are found in the years beginning with year +4, +5 or +6. The results thus suggest a significant 'rebound' into La Niña-like conditions in years +4, +5 and +6, as if the volcanic forcing would synchronize the internal clock of ENSO and pace a pattern of about one El Niño/La Niña-like alteration. As our significance estimates explicitly account for the serial correlation in the NINO3 series, this is not a simple consequence of the persistence structure of ENSO. Adjustment for post-eruption cooling (see Methods) slightly increases the prominence of post-event El Niño-like responses compared to subsequent La Niña-like responses, yielding the more physically intuitive result that the initial (El Niño) response is stronger than the subsequent (La Niña) rebound.

As an alternative evaluation of the volcano–El Niño connection independent of the SEA method, we determined (Table 1) the fraction of volcanic events followed in the first year by an El Niño-like response (see Methods). The results indicate that the likelihood of an El Niño event following an eruption in the subsequent cold season is significantly above that based on chance alone (which is roughly 25%) by generally falling into the ~35 to ~55% range (average is 42%) for the majority of the key-date lists. This represents nearly a doubling of the probability of an El Niño event. For 14 of the 20 lists, the increase in the likelihood of an El Niño event is significant at the $P < 0.05$ level ($P < 0.1$ for 16 of 20 cases). Although this latter analysis does not provide an ideal measure of the intrinsically multi-year nature of the inferred dynamical response, it does provide an independent, and perhaps more intuitive, measure of the strength of the volcano–El Niño relationship.

Our results do not indicate that explosive tropical eruptions trigger all El Niño events. Our analyses suggest, rather, that volcanic forcing drives the coupled ocean–atmosphere system more subtly towards a state in which multi-year El Niño-like conditions are favoured, followed by a weaker rebound into a La Niña-like state. This finding, though based on uncertain reconstructions of past ENSO behaviour, is entirely independent of previous analyses confined to the restricted instrumental climate record.

Although our analysis focuses on the system's response to explosive volcanism, our findings are nonetheless pertinent to understanding how the tropical Pacific ocean–atmosphere system may respond to exogenous (both natural and anthropogenic) radiative forcing on a variety of timescales. New evidence²⁹ suggests weaker interannual variability and a cooler mean state in the central tropical Pacific during the tenth–thirteenth centuries of the putative Medieval Warm Period, relative to an interval (seventeenth century) of the Little Ice Age. Given the present findings, such a trend would seem consistent with the response to the general increase in explosive volcanism during the fifteenth–nineteenth centuries in conjunction with reduced solar irradiance³⁰ that is responsible for the millennial cooling trend of Northern Hemisphere mean temperature before modern anthropogenic warming. The possibility of opposite forced changes in tropical and extratropical temperature variations in past centuries underscores the need for a more

complete understanding of the complex dynamical response of the climate to past forcing. □

Methods

The SEA method and its application

The SEA method isolates signals that are difficult to detect against relatively large background noise through application of composites^{21–24}. In contrast to Handler^{1,2}, we (1) use conventional and more robust, 'standardized' SEA methods developed to guard against statistical outliers, (2) accommodate a null hypotheses of equal pre- and post-eruption means and of post-eruption cooling, and (3) explicitly account for serial correlation in determining significance.

We apply the SEA to the reconstructed cold-season NINO3 and winter-mean SOI indices, using 20 'key-date' lists based on the IVI and VEI. Individual lists are selected using criteria with varying stringency with regard to the threshold eruption magnitude required for 'climatic relevance' and the temporal spacing of eruptions. For each key date in a list, a 21-yr window centred on the key date (defined as year 0) was extracted from the NINO3 or SOI index and entered into the SEA matrix. This provides a reasonable interval for resolving a response to volcanic forcing (~10 yr). This window constitutes the composite matrix's width dimension, whereas its length equals the number of key dates. Each composite matrix is centred by removal of the mean across its width.

Experiments

To independently reassess the volcano–El Niño hypothesis, experiments are performed during the 'pre-instrumental' period (that is, 1649–1868 for NINO3, and 1706–1868 for SOI). The reconstructions are also used for 'instrumental' analyses (1869–1979) (note that this period excludes El Chichón, an eruption pertinent to Handler's findings). Instrumental NINO3 and SOI assessments are performed for verification using the Handler key dates (1869–1982)^{1,2}. Of 20 lists, the pre-instrumental, instrumental, and 'full' (combined pre-instrumental/ instrumental) periods employ 7, 3 and 10 lists, respectively (Table 1). Given our theoretical physical mechanism and potential seasonal lags in response, we expect the strongest results in multi-year composites. Thus, in addition to the annual SEA, 3-yr composites are employed.

One criticism of the traditional SEA is its potential sensitivity to, for example, one particularly large El Niño event in the eruption matrix, leading to an apparently significant response^{3,4}. To guard against such bias, we also employ an alternative SEA in which anomaly values for each composite matrix row are normalized by the magnitude of the largest absolute anomaly magnitude in that row. This normalization reduces the potential leverage of any one large event. To further safeguard against outliers, we performed a ranked SEA procedure for the most vulnerable key-date lists. A ranked SEA is a version of SEA where the ENSO diagnostic anomalies extracted for each eruption are first ranked before insertion into the eruption matrix. The ranked SEA yields similar results to the traditional and normalized SEA (Table 1).

Data analysis

Global mean temperatures typically cool for several years following significant volcanic forcing³⁰. We thus use a null hypothesis of lower tropical SSTs (including the NINO3 area) following an eruption, independent of any dynamical (that is, ENSO-like) response. We invoke such a null hypothesis by removing statistically significant periods of post-eruption cooling before SEA initiation. Pre- and post-eruption periods for each list are tested for difference of means using the non-parametric Mann–Whitney test. If the composite post-eruption mean was significantly different from the pre-eruption composite mean ($P < 0.10$ using a one-sided test because post-eruption cooling is sought a priori), the pre- and post-eruption composites are separately mean-adjusted.

The Mann–Whitney results require adjustment in 10 of 20 reconstructed NINO3 composites (Supplementary Information). The mean cooling for these 10 lists is ~0.08 °C. However, a total of 18 out of 20 post-eruption means (not including the Handler list) are, on average, ~0.05 °C cooler than their corresponding pre-eruption means. Thus, the adjustment of only ten series is conservative.

Significance determination

The significance of composite responses is established using a Monte Carlo resampling procedure based on ensembles of 10,000 random surrogates derived from permutation of the composite matrices. Distributions for each composite year are generated from random reshuffling of anomalies for each eruption window (that is, row) and subsequent calculation of new 'random' composites. Each eruption year receives equal weight in distribution determination. We resample exclusively inside the eruption matrix, not the entire series. This preserves the eruption matrix statistics, but destroys any preferred temporal ordering between the pre- and post-eruption intervals.

The random 3-yr mean composite distributions are generated from the 10,000 random, annual composites. The resulting random 3-yr composite distributions are compared to the actual 3-yr composite responses to determine statistical significance.

We use a block resampling technique in determining significance in order to preserve the data's autocorrelation structure. From the actual eruption matrix, block resampling randomly generates an eruption matrix subsequently averaged to create a random composite matrix. Block resampling accounts for serial correlation when estimating null distributions from time series with autocorrelation. If omitted, the resulting confidence limits would be too liberal. The most straightforward block method is to resample the series in effectively independent blocks, rather than individual values. A first-order, autoregressive model was favoured by both Akaike and Bayesian information criteria, for which the effective correlation length is $\tau = -1/\log(\rho)$, where ρ is the lag-one

autocorrelation coefficient for the series ($\tau = 1$ for all cases in our study). The size of effectively independent blocks was therefore $2\tau = 2$.

We also calculated the fraction of times an eruption was followed, within one year, by an El Niño-like anomaly (defined as an anomaly of 0.3°C or greater—approximately 0.75 standard deviations for the long-term NINO3 reconstruction—relative to the post-eruption 10-yr mean). An El Niño 'fraction' was defined as the number of post-eruption El Niños normalized by the number of eruptions in each list. Significance was determined by a Monte Carlo resampling procedure analogous to that described above.

Received 20 February; accepted 1 November 2003; doi:10.1038/nature02101.

1. Handler, P. Possible association of stratospheric aerosols and El Niño type events. *Geophys. Res. Lett.* **11**, 1121–1124 (1984).
2. Handler, P. & Andsager, K. Possible association between the climatic effects of stratospheric aerosols and sea surface temperatures in the eastern tropical Pacific Ocean. *Int. J. Climatol.* **10**, 413–424 (1990).
3. Robock, A. Volcanic eruptions and climate. *Rev. Geophys.* **38**, 191–219 (2000).
4. Nicholls, N. Low-latitude volcanic eruptions and the El Niño/Southern Oscillation: A reply. *Int. J. Climatol.* **10**, 425–429 (1990).
5. Self, S., Rampino, M. R., Zhao, J. & Katz, M. G. Volcanic aerosol perturbations and strong El Niño events: No general correlation. *Geophys. Res. Lett.* **24**, 1247–1250 (1997).
6. Bjerknes, J. Atmospheric teleconnections from the equatorial Pacific. *Mon. Weath. Rev.* **97**, 163–172 (1969).
7. Zebiak, S. E. & Cane, M. A. A model El Niño/Southern Oscillation. *Mon. Weath. Rev.* **115**, 2262–2278 (1987).
8. Stahle, D. W. *et al.* Experimental dendroclimatic reconstruction of the Southern Oscillation. *Bull. Am. Meteorol. Soc.* **79**, 2137–2152 (1998).
9. Mann, M. E. *et al.* Global temperature patterns in past centuries: An interactive presentation. *Earth Interact.* **4–4**, 1–29 (2000).
10. Cane, M. A. *et al.* Twentieth-century sea surface temperature trends. *Science* **275**, 957–960 (1997).
11. Knutson, T., Manabe, S. & Gu, D. Simulated ENSO in a global coupled ocean-atmosphere model: Multidecadal amplitude modulation and CO_2 sensitivity. *J. Clim.* **10**, 138–161 (1997).
12. Meehl, G. A. & Washington, W. M. El Niño-like climate change in a model with increased atmospheric CO_2 concentrations. *Nature* **382**, 56–60 (1996).
13. Timmermann, A. *et al.* Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature* **398**, 694–697 (1999).
14. Boer, J. G., Flato, G., Reader, M. C. & Ramsden, D. A transient climate change simulation with greenhouse gas and aerosol forcing: Experimental design and comparison with the instrumental record for the 20th century. *Clim. Dyn.* **16**, 405–425 (2000).
15. Meehl, G. A. *et al.* Response of the NCAR climate system model to increased CO_2 and the role of physical processes. *J. Clim.* **13**, 1879–1898 (2000).
16. Noda, A., Yamaguchi, K., Yamaki, S., Yukimoto, S. Relationship between natural variability and CO_2 -induced warming pattern: MRI AOGCM experiment. *10th Symp. on Global Change Studies 10–15 January 1999, Dallas, TX.* (American Meteorological Society, 1999).
17. Clement, A. C., Seager, R., Cane, M. A. & Zebiak, S. E. An ocean dynamical thermostat. *J. Clim.* **9**, 2190–2196 (1996).
18. Hirono, M. On the trigger of El Niño Southern Oscillation by the forcing of early El Chichón volcanic aerosols. *J. Geophys. Res.* **93**, 5365–5384 (1988).
19. Robock, A. *et al.* GCM evaluation of a mechanism for El Niño triggering by the El Chichón ash cloud. *Geophys. Res. Lett.* **22**, 2369–2372 (1995).
20. Zielinski, G. A. Use of paleo-records in determining variability within the volcanism-climate system. *Quat. Sci. Rev.* **19**, 417–438 (2000).
21. Panofsky, H. A. & Brier, G. W. *Some Applications of Statistics to Meteorology* 159–161 (Penn. State University, University Park, Pennsylvania, 1958).
22. Bradley, R. S., Diaz, H. F., Kiladis, G. N. & Eischeid, J. K. ENSO signal in continental temperature and precipitation records. *Nature* **327**, 497–501 (1987).
23. Seager, C. B., Kelly, P. M., Jones, P. D. & Goodess, C. M. Global surface-temperature responses to major volcanic eruptions. *Nature* **330**, 365–367 (1987).
24. Haurwitz, M. W. & Brier, G. W. A critique of superposed epoch analysis method: Its application to solar-weather relations. *Mon. Weath. Rev.* **109**, 2074–2079 (1981).
25. Simpkin, T. & Siebert, L. *Volcanoes of the World* 2nd edn (Smithsonian Institution, Geoscience Press, Tucson, Arizona, 1994).
26. Robock, A. & Free, M. P. Ice cores as an index of global volcanism from 1850 to the present. *J. Geophys. Res.* **100**, 11549–11567 (1995).
27. Können, G. P., Jones, P. D., Latofen, M. H. & Allan, R. J. Pre-1866 extensions of the Southern Oscillation Index using early Indonesian and Tahitian meteorological readings. *J. Clim.* **11**, 2325–2339 (1998).
28. Kaplan, A., Cane, M. A., Kushnir, Y. & Clement, A. C. Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.* **103**, 18567–18589 (1998).
29. Cobb, K. M., Charles, C. D., Edwards, R. L., Cheng, H. & Kastner, M. El Niño-Southern Oscillation and tropical Pacific climate during the last millennium. *Nature* **424**, 271–276 (2003).
30. Crowley, T. J. Causes of climate change over the past 1000 years. *Science* **289**, 270–277 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This research was supported (M.E.M. and J.B.A.) by the NOAA- and NSF-supported 'Earth Systems History' programme and the NCAR-Early Career Scientist Assembly (C.M.A.). We thank R. S. Bradley for his support of C.M.A. during early stages of this work, R. E. Davis for suggestions with respect to the statistical procedure, and A. Robock and K. Trenberth for comments on the manuscript. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.B.A. (jba7g@virginia.edu).

A newly discovered species of living baleen whale

Shiro Wada¹, Masayuki Oishi² & Tadasu K. Yamada³

¹National Research Institute of Fisheries Science, Fisheries Research Agency, 2-12-4 Fukuura, Kanazawa-ku, Yokohama, 236-8648, Japan

²Iwate Prefectural Museum, 34 Ueda-Matsuyashiki, Morioka, 020-0102, Japan

³National Science Museum, 3-23-1 Hyakunin-cho, Shinjuku-ku, Tokyo, 169-0073, Japan

In the late 1970s eight *Balaenoptera* specimens of unknown identity were caught in the lower latitudinal Indo-Pacific waters by Japanese research whaling vessels¹. The combination of the allozyme patterns and physical maturity of the eight specimens separated them from all acknowledged *Balaenoptera* species². In September 1998 we collected a medium-sized baleen whale carcass on a coastal island in the Sea of Japan. This specimen and the previously collected eight specimens resembled *Balaenoptera physalus* (fin whale) in external appearance but were much smaller. Comparison of external morphology, osteology and mitochondrial DNA data grouped the nine specimens as a single species but separated them from all known baleen whale species. Therefore, here we describe a new species of *Balaenoptera*, which is characterized by its unique cranial morphology, its small number of baleen plates, and by its distant molecular relationships with all of its congeners. Our analyses also separated *Balaenoptera brydei* (Bryde's whale)^{3,4} and *Balaenoptera edeni* (Eden's whale)⁵ into two distinct species, raising the number of known living *Balaenoptera* species to eight.

Cetacea Brisson, 1762

Mysticeti Flower, 1864

Balaenopteridae Gray, 1864

Balaenoptera Lacépède, 1804

Balaenoptera omurai sp. nov.

Etymology. The specific name is in honour of the late H. Omura, a Japanese cetologist, for his contribution to the knowledge of Cetacea.

Holotype. Adult female, NSMT-M32505, National Science Museum, Tokyo. A complete skeleton, both complete baleen rows and frozen pieces of muscle, blubber and kidney were collected at Tsunoshima Island (34° 21' 03" N, 130° 53' 09" E) by T.K.Y., T. Kuramochi, M.O., E. Jibiki and S. Fujioka. The collection was undertaken three days after the accidental death of the animal.

Paratypes. Five females and three males, NRIFS1–8, National Research Institute of Far Seas Fisheries, Fisheries Research Agency, Shizuoka (see Table 1). The longest baleen plate, an earplug and a piece of the sixth thoracic vertebra with epiphysis were collected by trained staff from each animal. NRIFS6 includes 18 more baleen plates.

Locality. The Sea of Japan (type locality), the Solomon Sea and the eastern Indian Ocean near the Cocos Islands.

Diagnosis. *Balaenoptera omurai* differs from all of its congeners by having the following unique characters: medially expanded posterior portion of ascending process of maxilla, which conceals posterior end of premaxilla along the adjacent nasal (Fig. 1a, f); approximately 200 baleen plates on one side, which is the smallest number among all of its congeners except for *B. edeni* (the baleen plate number of which is still unknown); 21 diagnostic sites in the complete mitochondrial (mt)DNA control region sequence.

Discussion. The body length is less than 12 m (Table 1), similar to *B. edeni*^{5,6}. The skull is relatively broad and flat (Fig. 1a–c). The rostrum is tapering from its base, and its lateral margin is

conspicuously convex in dorsal view. Maximum width between the dorsomedial rims of the maxillae is at the lowest range of those for balaenopterids, occupying approximately 30% of the basal width of the rostrum. The ascending process of the maxilla broadens towards its posterior end (Fig. 1f), whereas in *B. edeni* it is slender throughout its length, giving a large space for the nasals and premaxillae (Fig. 1g, h). The frontal is only exposed as a narrow belt between the ascending processes of the maxillae and the supraoccipital, and between the nasals and a small exposure of the interparietal (Fig. 1f), whereas in *B. edeni* the frontal is exposed broadly posterior to and between the ascending processes of the maxillae, and forms a low but noticeable protuberance, which looks like a pedestal for each ascending process of the maxilla (Fig. 1g, h). The posterior end of the ascending process of the maxilla and the anterior margin of the supraoccipital are lying between levels of the antorbital and postorbital processes of the frontal. The posteromedial part of the supraorbital process of the frontal is broadly covered by the parietal, semicircularly when viewed dorsally (Fig. 1a). The hamular process of the pterygoid is broad and short, situated around the level of the posterior rim of the subtemporal fossa. The alisphenoid is separate to the squamosal (Fig. 1e), whereas in *B. edeni* it has a contact with the squamosal. The dorsal portion of the bifurcated pterygoid process of the squamosal is separate to the palatine. Two small foramina open along the suture between the parietal and squamosal in the wall of the braincase. The foramen pseudo-ovale surrounded only by the squamosal opens just posterior to the medial portion of the crest between the subtemporal and glenoid fossae, whereas in *B. edeni* and *B. brydei* the foramen pseudo-ovale surrounded by the squamosal and pterygoid opens in the glenoid fossa. Unlike *Balaenoptera musculus* (blue whale), *B. physalus* and *Balaenoptera borealis* (sei whale), the angle of the mandible conspicuously projects posteriorly, ending after the posterior edge of the condyle (Fig. 1d). The head of the first rib is not bifurcated, and its distal part is not broadened laterally like a blade. All epiphyses of the vertebrae except for NRIFS8 are fused to each centrum. The vertebral and phalangeal formulae are given as cervical (7) + thoracic (13) + lumbar (12) + caudal (21) = 53 and I-5, II-7, IV-6, V-3, apart from four carpals and four metacarpals, respectively.

The external appearance of *B. omurai* resembles that of *B. physalus* in the following respects: the left side of the throat is widely pigmented and the remaining ventral surface is uniformly white, giving an asymmetrical appearance (Fig. 2a); the anterior edge and inner surface of both flippers are white from tip to shoulder; the ventral surface of the tail flukes is white with a black margin; the

ventral grooves extend behind the umbilicus and the total number of grooves is estimated as 80 to 90. The left gape is white in NRIFS1 (no observation was made on the right gape). Although the surface of the head is not entirely smooth, the lateral ridges⁷ are absent, which are prominent in *B. brydei* and probably in *B. edeni*.

The baleen plates are, however, quite different compared with those of *B. physalus* in size, shape, colour and number. They are short and broad with uncured, stiff, greyish-white fringes. The dimensions of the longest baleen plates ($n = 9$) are 26.0 ± 2.4 cm (mean $\pm$ s.d.) in length and 21.4 ± 1.9 cm in breadth. The ratio of length to breadth, 1.22 ± 0.15 , is considerably smaller than the value of 1.48 ± 0.11 ($n = 7$) for the Antarctic *B. physalus*⁸. Baleen colour is just like that of *Balaenoptera bonaerensis* (Antarctic minke whale)^{9,10}, varying with the position in the baleen row, and displays asymmetry: in the right row of NSMT-M32505 the anterior 61 plates are all yellowish-white in colour, the posterior 40 plates are all black, and the intermediate 102 plates are bi-coloured (black externally and yellowish-white internally) (Fig. 2b); whereas in the left row of the same specimen, the anterior 178 plates are bi-coloured and the remaining 30 plates are all black; in both rows, the proportional width of the yellowish-white band becomes gradually smaller in the posterior plates. Similar transitional change of baleen colour is also seen in NRIFS6. The proportion of the yellowish-white band to the basal width of the longest baleen plate is $30.7 \pm 6.2\%$ ($n = 5$) in the right plates and $13.7 \pm 7.3\%$ ($n = 4$) in the left plates, showing that the yellowish-white band is dominant in the right baleen row ($P < 0.001$) as in *B. physalus*^{8,11}, *B. borealis*¹², *B. bonaerensis*^{9,10} and the diminutive form of *Balaenoptera acutorostrata* (common minke whale)^{9,10}. The total number of baleen plates on one side—203 (right) and 208 (left) counted in NSMT-M32505 and 181–190 (right) estimated for NRIFS6—is markedly smaller than that of the other *Balaenoptera* species.

The phylogenetic position of *B. omurai* relative to its congeners was examined using the complete control region sequence of the mtDNA molecule. The length of the control region in the three *B. omurai* specimens—one from the Sea of Japan (NSMT-M32505; GenBank accession number AB116095), one from the Solomon Sea (NRIFS1; AB116096) and one from the eastern Indian Ocean (NRIFS7; AB116097)—was 938 nucleotides, and they only differed by 4–5 nucleotides. The number of nucleotide differences between *B. omurai* and its congeners in 901 aligned sites (gaps excluded) was 62–67 for *B. brydei* (77N62 (ref. 13); AB116098), 65–70 for *B. edeni* (RMNH4003 (ref. 6); AB116099), 66–71 for *B. borealis*, 70–73 for *B. musculus*, 73–78 for *B. physalus*, 89–94 for *B. acutorostrata* and 92–97 for *B. bonaerensis*. The smallest interspecific nucleotide

Table 1 Biological information on the nine specimens of *B. omurai* sp. nov.

Museum number	Collected baleen plates	Body length (m)	Sex	Ovaries‡ or testes§	Age	Physical maturity	Date of catch/stranding	Position	
								Lat.	Long.
NRIFS1*	1 (R)	11.5	Female	1-11	29	a	24/10/1976	10° 03' S,	157° 29' E
NRIFS2	1 (L)	9.6	Male	1.9/1.7	38	a	24/10/1976	9° 53' S,	157° 37' E
NRIFS3	1 (L)	11.2	Female	0-6	9	A	24/10/1976	9° 57' S,	157° 41' E
NRIFS4	1 (L)	10.0	Male	1.6/1.5	21	A	24/10/1976	9° 49' S,	157° 29' E
NRIFS5	1 (L)	10.3	Female	1-12	23	A	24/10/1976	10° 07' S,	157° 51' E
NRIFS6	19 (R)†	9.6	Male	2.2/2.1	34	A	24/10/1976	10° 17' S,	157° 56' E
NRIFS7	1 (R)	10.4	Female	0-4	19	A	15/11/1978	10° 51' S,	97° 02' E
NRIFS8	1 (R)	10.1	Female	0-3	18	n	17/11/1978	10° 53' S,	94° 29' E
NSMT-M32505	All (L&R)	11.03	Female	NC	NC	A	11/09/1998	34° 21' N,	130° 50' E

Age (column six) is indicated as the number of growth layers in an earplug. Physical maturity (column seven) was determined based on the extent of epiphyseal fusion in the sixth thoracic vertebra using the criteria: A, fused, no sign of joint; a, fused, joint visible; n, not fused, with thin cartilage. 'A' and 'a' represent physically mature whales whereas 'n' represents physically immature whale.

*Japanese inspectors took 8-mm movie footage and photographs. The description of the external appearance of the new species is based on these pictures and our direct observation on NSMT-M32505.

The throat colour asymmetry is shown by this specimen only.

†Nineteen baleen plates from the right (R) row were collected at every tenth position, starting from the first one and excluding the rudimentary hairs; that is, the first, the eleventh, the twenty-first, and so on. The thirteenth plate was the longest. A total number of 181–190 baleen plates is estimated for the right row.

‡The left value indicates the number of corpora lutea in both ovaries, and the right value the combined number of corpora lutea and corpora albicantia. NC, not collected. Females with one or more corpora are sexually mature. None of the females were pregnant.

§Shown in kilograms.

|| Damaged or partially lost.

difference, 31 nucleotides, was between *B. borealis* and *B. brydei*. Thus, the difference between *B. omurai* and any of its congeners is much greater than that between these acknowledged species. Note that the difference between *B. edeni* and *B. brydei*, 35 nucleotides, is slightly greater than that between *B. borealis* and *B. brydei*, implying that they are also distinct species. The separate position of *B. omurai* in the *Balaenoptera* tree (Fig. 3) was conclusively supported irrespective of whether or not the highly variable portion (first 173 nucleotides)¹⁴ of the control region was included in the analyses.

Balaenopterid nomenclature is currently confusing¹⁵. The con-

fusion is at least partly related to the question of whether Bryde's whale is one species (*B. edeni*)⁶ or two (*B. edeni* and *B. brydei*)^{16–18}, and to the recent assertion that our eight specimens with anomalous allozymes may belong to *B. edeni*^{19,20}. However, the current molecular analyses have conclusively separated *B. omurai* from the *borealis/brydei/edeni* group, and separated *B. edeni* from the *borealis/brydei* group. Furthermore, we have confirmed that the holotype (GRM223 (ref. 5), Indian Museum, Kolkata), Sidhi Island (Indian Museum)²¹ and Pul(a)u Sugi (RMNH4003, National Museum of Natural History, Leiden) *B. edeni* skulls share the previously mentioned unique characters: the slender ascending

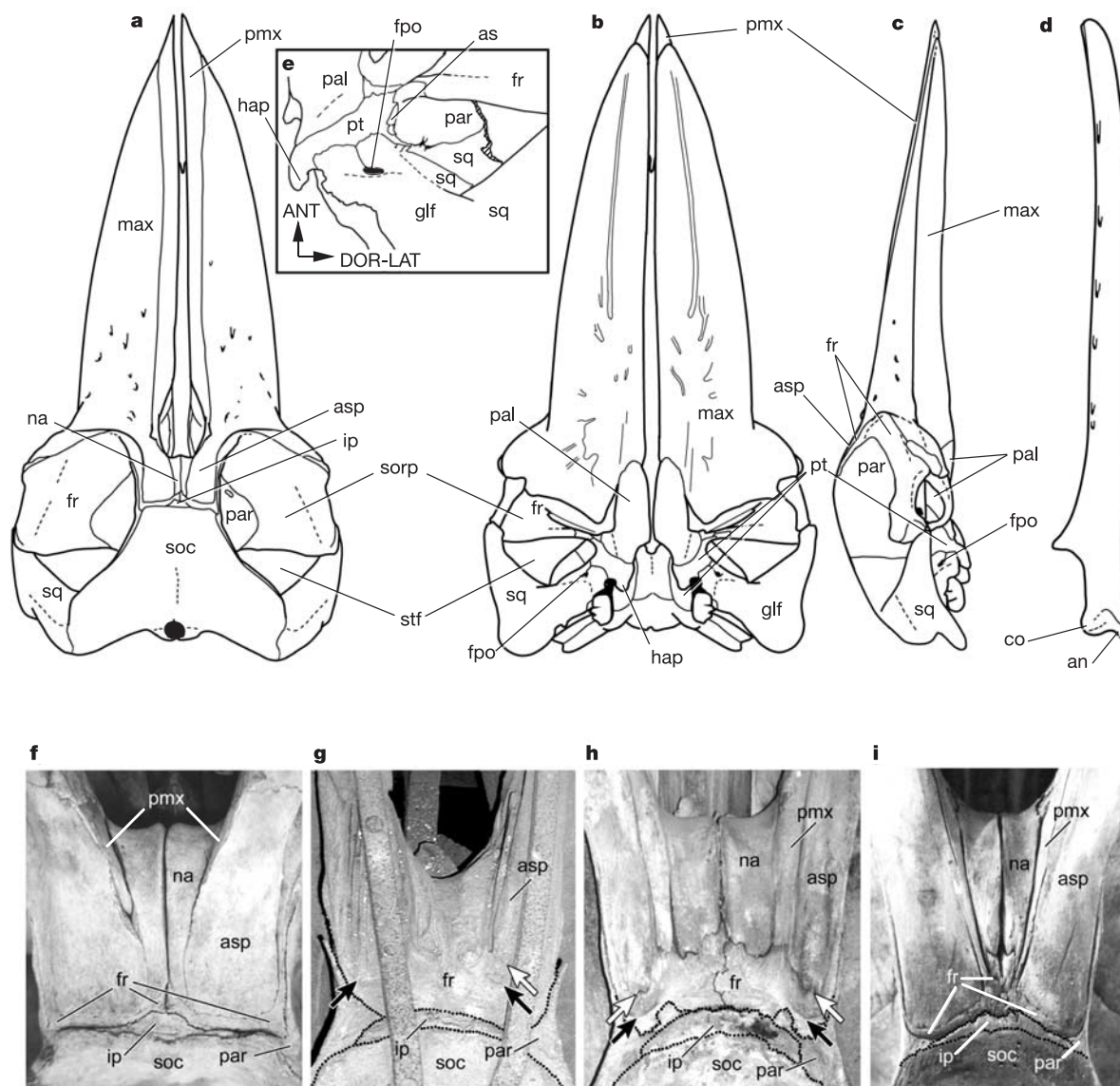


Figure 1 The holotype skull of *B. omurai* sp. nov. (NSMT-M32505). **a–d**, The skull is shown in dorsal (**a**), ventral (**b**) and lateral (**c**) views, and the mandible (NSMT-M32505) is shown in lateral and slightly oblique view (**d**) to the median plane of the skull. Length and width of the skull and length of the mandible are 2,832 mm, 1,454 mm and 2,790 mm, respectively. **e**, Ventrolateral view around the left subtemporal fossa. **f–i**, Comparison of the dorsal view of the skull at vertex among the three species. **f**, NSMT-M32505 (*B. omurai* holotype), medial margin of the left ascending process of the maxilla is broken and the posterior end of the premaxilla is exposed. **g**, **h**, The posterior end of the protuberance of the frontal (black arrow) is located a few centimetres posterior to the

posterior end of the ascending process (white arrow). **g**, GRM223 (*B. edeni* holotype)⁵, both of the nasals and the premaxillae as well as the left ascending process are lost. **h**, RMNH4003 (*B. edeni*)⁶. **i**, 78N33 (*B. brydei*, adult female)¹³. an, angle; ANT, anterior; as, alisphenoid; asp, ascending process; co, condyle; DOR-LAT, dorsolateral; fpo, foramen pseudo-ovale; fr, frontal; glf, glenoid fossa; hap, hamular process; ip, interparietal; max, maxilla; na, nasal; pal, palatine; par, parietal; pmx, premaxilla; pt, pterygoid; soc, supraoccipital; sorp, supraorbital process; sq, squamosal; stf, subtemporal fossa.

process of the maxilla and the broadly exposed frontal with the pedestal-like protuberance (Fig. 1g, h). These features are absent in all other *Balaenoptera* species—in *B. brydei* the ascending process of the maxilla is broad and the frontal has a small exposure without protuberance (Fig. 1i). These osteological findings, coupled with

molecular ones, corroborate our interpretation that *B. brydei*, *B. edeni* and *B. omurai* are distinct species. It is expected that, based on these diagnostic characters, more specimens of *B. omurai* and *B. edeni* will be reported from the sea as well as from museum collections. □

Received 12 June; accepted 25 September 2003; doi:10.1038/nature02103.

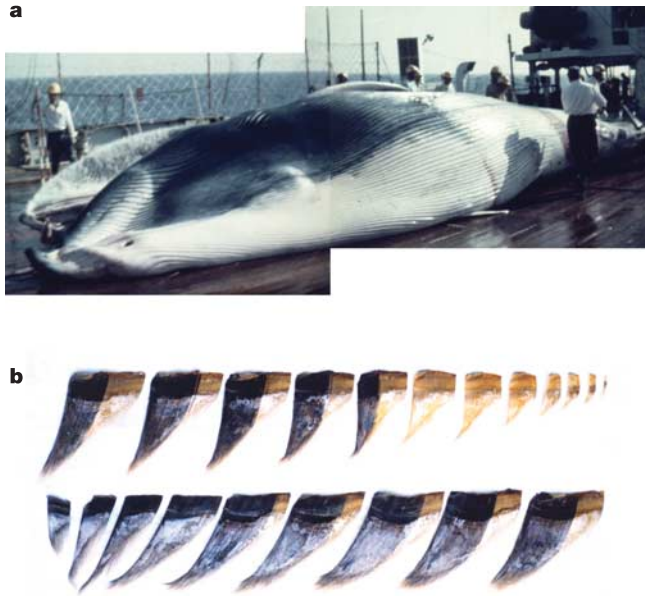


Figure 2 External morphology of *B. omurai* sp. nov. **a**, Ventral view of NRIFS1. **b**, Anterior view of 21 baleen plates (NSMT-M32505) collected from every tenth position starting from the first one of the right row.

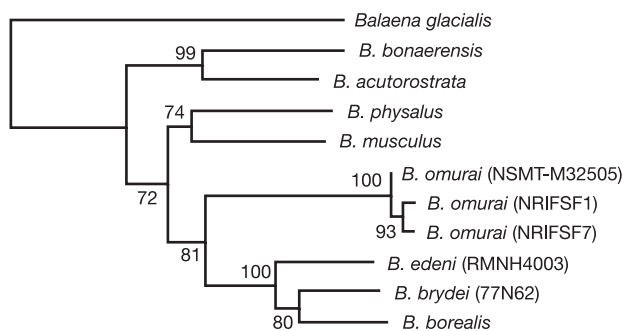


Figure 3 Neighbour-joining tree²² of the complete mtDNA control region sequences of all eight species suggested to belong to *Balaenoptera*, constructed using Kimura's two-parameter model in MEGA²³ with *Balaena glacialis* (right whale) as the outgroup. Fifty-three sites with one or more gaps were removed from the original alignment obtained by using Clustal W²⁴, resulting in a 901-nucleotide-long final alignment. Values at each node represent bootstrap values based on 1,000 replicates. The following six sequences were downloaded from GenBank: M60408 (*B. bonaerensis*)²⁵, X61145 (*B. physalus*)²⁶, X72195 (*B. borealis*)¹⁴, X72199 (*Balaena glacialis*)¹⁴, X72204 (*B. musculus*)²⁷ and X72006 (*B. acutorostrata*)¹⁴. The remaining five sequences were amplified by polymerase chain reaction using a primer set designed from X61145 and X72204—5'-ACACCTCC TAAGACTCAAGGAAG-3' (on tRNA-(Thr/Pro)) and 5'-TAGACATTTTCAGTGTCTTGTCTT-3' (on tRNA-(Phe)). Template DNA was extracted from muscle in ethanol (NSMT-M32505), baleen plate (NRIFS1), dried soft tissue (RMNH4003) and frozen liver (NRIFS7 and 77N62 (ref. 13)). The amplified products were cloned in pGEM-T vector and sequenced for both strands.

- Ohsumi, S. Population study of the Bryde's whales in the Southern Hemisphere under scientific permit in the three seasons, 1976/77–1978/79. *Rep. Int. Whal. Commis.* **30**, 319–331 (1980).
- Wada, S. & Numachi, K. Allozyme analyses of genetic differentiation among the populations and species of the *Balaenoptera*. *Rep. Int. Whal. Commis.* **13** (special issue), 125–154 (1991).
- Olsen, Ö. On the external characters and biology of Bryde's whale (*Balaenoptera brydei*), a new orca from the coast of South Africa. *Proc. Zool. Soc. Lond.*, 1073–1090 (1913).
- Lönnberg, E. The skeleton of *Balaenoptera brydei* Ö. Olsen. *Arkiv. f. Zoologi.* **23A1**, 1–23 (1931).
- Anderson, J. in *Anatomical and Zoological Researches* Vol. 1, 551–564 (B. Quaritch, London, 1878/9).
- Junge, G. C. A. On a specimen of the rare fin whale, *Balaenoptera edeni* Anderson stranded on Pulu Sugi near Singapore. *Zoologische Verhandlungen* **9**, 1–26 (1950).
- Omura, H. in *Whales, Dolphins and Porpoises* (ed. Norris, K. S.) 70–78 (Univ. California Press, Berkeley/Los Angeles, 1966).
- Mackintosh, N. A. & Wheeler, J. F. G. Southern blue and fin whales. *Discov. Rep.* **1**, 257–540 (1929).
- Best, P. B. External characters of southern minke whales and the existence of a diminutive form. *Sci. Rep. Whales Res. Inst.* **36**, 1–33 (1985).
- Arnold, P., Marsh, H. & Heinsohn, G. The occurrence of two forms of minke whales in east Australian waters with description of external characters and skeleton of the diminutive form. *Sci. Rep. Whales Res. Inst.* **38**, 1–46 (1987).
- True, F. W. The whalebone whales of the western north Atlantic. *Smithson. Contrib. Knowl.* **33**, 1–332 (1904).
- Matthews, L. H. The sei whale, *Balaenoptera borealis*. *Discov. Rep.* **17**, 183–290 (1938).
- Omura, H., Kasuya, T., Kato, H. & Wada, S. Osteological study of the Bryde's whale from the central south Pacific and eastern Indian Ocean. *Sci. Rep. Whales Res. Inst.* **33**, 1–26 (1981).
- Árnason, Ú., Gullberg, A. & Widegren, B. Cetacean mitochondrial DNA control region: Sequences of all extant baleen whales and two sperm whale species. *Mol. Biol. Evol.* **10**, 960–970 (1993).
- Rice, D. W. Marine mammals of the world: Systematics and distribution. *Mar. Mamm. Sci.* **4** (special publication), 67–78 (1998).
- Soot-Ryen, T. On a Bryde's whale stranded on Curaçao. *Norsk Hvalfangstid* **50**, 323–332 (1961).
- Pilleri, G. & Gihl, M. in *Investigations on Cetacea* Vol. 5 (ed. Pilleri, G.) 95–149 (Univ. Berne, Berne, 1973/4).
- Best, P. B. Two allopatric forms of Bryde's whale off South Africa. *Rep. Int. Whal. Commis.* **1** (special issue), 10–38 (1977).
- LeDuc, R. G. & Dixon, A. E. in *Molecular and Cell Biology of Marine Mammals* (ed. Pfeiffer, C. J.) 100–110 (Krieger, Florida, 2002).
- Kato, H. in *Encyclopedia of Marine Mammals* (eds Perrin, W. F., Würsig, B. & Thewissen, J. G. M.) 171–177 (Academic, San Diego, 2002).
- Andrews, R. C. A note on the skeleton of *Balaenoptera edeni*, Anderson, in the Indian Museum, Calcutta. *Rec. Indian Mus.* **15**, 105–107 (1918).
- Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
- Kumar, S., Tamura, K. & Nei, M. MEGA: Molecular Evolutionary Genetics Analysis V.1.01 (Pennsylvania State Univ., Univ. Park, 1993).
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
- Hoelzel, A. R., Hancock, J. M. & Dover, G. A. Evolution of cetacean mitochondrial D-loop region. *Mol. Biol. Evol.* **8**, 475–493 (1991).
- Árnason, Ú., Gullberg, A. & Widegren, B. The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. *J. Mol. Evol.* **33**, 556–568 (1991).
- Árnason, Ú. & Gullberg, A. Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. *J. Mol. Evol.* **37**, 312–322 (1993).

Acknowledgements We thank C. Smeenk for comments, providing *B. edeni* tissue sample, paperwork for the CITES permit and critical reading of the manuscript; K. Ikemoto, S. Shimizu and T. Yokomachi for photographs of NSMT-M32505 and NRIFS1; J. R. B. Alfred, S. Biswas, H. P. Castello, R. Chakravarti, S. K. Chakravarti, C. Dechsakulwatana, B. Fernholm, R. Harbord, P. D. Jenkins, Y. Musig, Y. Newi, S. K. Podder, C. W. Potter, P. Rudd, G. C. Ray, S. Said, A. K. Sanyal, R. Sabin and O. B. Vaccaro for permission to examine the specimens in their custody; Ú. Árnason, C. C. Kinze and J. G. Mead for comments and critical reading of the manuscript; L. B. Holthuis, K. Hosoya, Y. Masaki, M. C. Milinkovitch, N. Miyazaki and K. Numachi for comments; H. Hatanaka, S. Ohsumi and Y. Shimadzu for permission to use the data and materials from the research whaling; M. Etoh, S. Fujioka, M. Furuno, A. Tojima and T. Uchiyama for logistic support and help with preserving the holotype; T. Kakuda and T. Kobayashi for help with the phylogenetic analysis; T. Imamura for help with the research at Indian Museum; H. Ishikawa, T. Itoh, E. Jibiki, K. Kuramochi, K. Nakamura, T. Shimokawa and K. Tokutake for assistance in the field; and K. Arai, R. Imai, N. Kaneko, K. Maeda, M. Ogino, K. Shibayama, Y. Tajima, A. Umetani and Y. Watanabe for sample preparation and data collection.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W. (wadash@affrc.go.jp) or T.K.Y. (yamada@kahaku.go.jp). The DNA sequences reported in this study have been deposited in GenBank under accession numbers AB116095–AB116099.

Compartments revealed in food-web structure

Ann E. Krause¹, Kenneth A. Frank^{1,2}, Doran M. Mason³, Robert E. Ulanowicz⁴ & William W. Taylor¹

¹Department of Fisheries and Wildlife, and ²Department of Counseling, Educational Psychology, and Special Education, Michigan State University, East Lansing, Michigan 48824, USA

³NOAA, Great Lakes Environmental Research Laboratory, 2205 Commonwealth Boulevard, Ann Arbor, Michigan 48105, USA

⁴Chesapeake Biological Laboratory, University of Maryland, Solomons, Maryland 20688-0038, USA

Compartments¹ in food webs are subgroups of taxa in which many strong interactions occur within the subgroups and few weak interactions occur between the subgroups². Theoretically, compartments increase the stability in networks^{1–5}, such as food webs. Compartments have been difficult to detect in empirical food webs because of incompatible approaches^{6–9} or insufficient methodological rigour^{8,10,11}. Here we show that a method for detecting compartments from the social networking science^{12–14} identified significant compartments in three of five complex, empirical food webs. Detection of compartments was influenced by food web resolution, such as interactions with weights. Because the method identifies compartmental boundaries in which interactions are concentrated, it is compatible with the definition of compartments. The method is rigorous because it maximizes an explicit function, identifies the number of non-overlapping compartments, assigns membership to compartments, and tests the statistical significance of the results^{12–14}. A graphical presentation¹⁴ reveals systemic relationships and taxa-specific positions as structured by compartments. From this graphic, we explore two scenarios of disturbance to develop a hypothesis for testing how compartmentalized interactions increase stability in food webs^{15–17}.

Similarity between human social networks and food webs has recently been noted through exchanges between ecologists and sociologists^{18,19}. In the social sciences, cohesive subgroups in human communities have been an important concept since the 1950s, when it was proposed that social systems were more efficient and durable when composed of subgroups in which interactions were concentrated^{3,20,21}. The concept of cohesive subgroups has had strong theoretical support, but the methodologies needed to apply the concept to communities were lacking^{12,13}. This is also the case for food-web compartments in ecology, wherein methods for identifying compartments have often emphasized the similarity of prey and predators between taxa^{6–9}, which results in little direct interaction or carbon exchange within compartments.

Here we use a recently developed social network method^{12,14} for identifying cohesive subgroups to detect compartments in empirical food webs. The method identifies compartments in which interactions are concentrated, thus conserving the flow of energy and organic material, for example, within compartments, just as information and influence flow primarily within human subgroups in which interactions are concentrated. Although the algorithm identifies compartments in which interactions are concentrated, interactions are not exclusively confined within compartments. Thus there are critical cross-compartment interactions that integrate the compartments into a food web¹, just as interactions between people in different subgroups sustain social systems and societies^{3,20,21}.

Crucial to the network method is the criterion defining the concentration of interactions within compartments, where an interaction is predator taxon *i* consuming prey taxon *i'*. The

criterion is the increase in the odds of an interaction occurring (as opposed to not occurring) given that two taxa are in the same compartment as opposed to in different compartments. This odds ratio corresponds to $[(A \times D)/(B \times C)]$ as defined in Table 1 and can be interpreted as a comparison of density of interactions within versus between compartments and is thus associated with key parameters for social network models¹². The statistical significance of the odds ratio is determined by Monte Carlo simulations (see Methods). In the food-web literature²², density is analogous to interactive connectance (IC), with overall IC defined by $[(B+D)/(A+B+C+D)]$ and interpreted as the proportion of realized interactions out of all possible interactions.

The algorithm we employed has four features^{12–14} that are crucial for our application^{6,23}. First, taxa are assigned to non-overlapping compartments as a result of the odds ratio being a flexible criterion^{12–14}. Thus, the assignment of each taxon contributes to the concentration of interactions within all of the compartments of the food web¹². Second, the algorithm generally does not require a priori specification of the number of compartments. This feature removes a key subjective decision from the procedure and assists in simulating a sampling distribution for the odds ratio. Third, the algorithm was calibrated by applying it to extensive simulated data with known compartment assignments^{12,13}. Fourth, compartment boundaries can be embedded in a graphical presentation of the food web¹³, thus facilitating the interpretation of the roles of compartments and their taxa within food webs. None of the previous methodologies in either social networks^{11–14} or ecology^{6–11} have all four of these important features.

We applied this method to five food webs: Ythan Estuary²⁴, Little Rock Lake^{22,25}, St Martin Island²⁶, Chesapeake Bay^{27,28} and a cypress wetland (See Supplementary Information A). Seventeen separate versions of these food webs with various levels of aggregation of taxa, weight of interactions and season were considered. Seven of them yielded odds ratios that were statistically significantly greater than would be expected by chance alone (Table 2; $\alpha = 0.05$). Six of these would still be statistically significant when adjusting for the number of tests by using a Bonferroni correction of α . As we expected, IC within compartments was higher and IC between compartments was lower than the overall IC (by factors of 1.9 to 3.5 and 0.003 to 0.27 respectively). The algorithm did not detect compartments in St Martin Island, a narrow food web focused on two lizard species and not likely to have compartments, but did detect compartments in broader food webs (for example, Chesapeake Bay). This is consistent with our claim that the method is able to detect the presence or absence of compartments with reasonable accuracy.

The resolution of a food web can affect the detection of compartments. We detected compartmentalization in only 1 of 14 less complex food webs originally analysed^{6,8} (Supplementary Information B) as opposed to three of five more complex food webs presented here. No compartments were detected in the aggregated version of Little Rock Lake or in the unweighted versions of Chesapeake Bay and the cypress wetland, whereas compartments were detected in alternative versions. These inconsistencies suggest that ignoring weights when aggregating taxa decreases the number of analysed interactions and can obscure strong relationships that contribute to compartmentalization. Note that the range in weights for interactions in all of our food-web versions was at the upper limit for the method. Ideally, the range should be narrower (for example, from 1 to 100). A wide range in weights (for example, from 1 to 99,999) can result from high aggregation in the basal taxa and low aggregation in top predators, such as in our weighted food webs.

The graphical presentation (Fig. 1) shows the compartmental structure of the Chesapeake Bay food web with 45 taxa and weighted by interaction strength. Although the detail in this graphic seems complex²³, the image reveals an intuitive understanding of the food

Table 1 Association between compartment membership and occurrence of interactions between taxa

Compartment membership	Different Same	Interaction occurring		$n(n-1) - \Sigma_g n_g(n_g-1)$ $\Sigma_g n_g(n_g-1)$
		No	Yes	
		A C	B D	
		$n(n-1) - \Sigma_i \Sigma_{i'} X_{ii'}$	$\Sigma_i \Sigma_{i'} X_{ii'}$	$n(n-1)$

$X_{ii'}$ represents the presence (1) or absence (0) of an interaction between taxon i and taxon i' , n the number of taxa in the food web, and n_g the number of taxa in compartment g . Interactions between predator taxon i and prey taxon i' can be weighted by integers from 0 to W_m , the maximum weight across all ii' interactions. Weights are included in the above table by multiplying $X_{ii'}$ by the weight assigned to the ii' interaction and by multiplying $n(n-1)$ and $\Sigma_g n_g(n_g-1)$ by W_m (ref. 13). A represents unrealized interactions between compartments, B realized interactions between compartments, C unrealized interactions within compartments, and D realized interactions within compartments.

web. Compartment A has 28 taxa, most of which would be considered pelagic (in the water column) and compartment B has 17 taxa that are primarily benthic (in sediments). Some taxa placements might seem counterintuitive when considering the physical component of habitat. For example, clams (27, 28 and 29) and oyster (30) physically reside in the benthos, like the clams (25 and 26) in compartment B. Our results demonstrate that the biotic habitat of 27, 28, 29 and 30 is in the pelagia because of their strong interactions with bacteria (3, 4 and 5) and ciliates (7, 8 and 9) in compartment A, which supports previous research²⁷. Compartments should measure biotic habitat^{8,10,11}, and compartment membership from our other significant results (Supplementary Information C) lend additional support. At the system level, compartments A and B are linked through cross-compartment interactions. The few weak interactions indicate the level of isolation between the two compartments. The interactions that compartment B has with A are more evenly dispersed within its compartment, where 65% of its taxa interact with A. Conversely, only 30% of the taxa in compartment A interact with B and these interactions are concentrated within specific areas of A.

Placement of a taxon indicates its role within its compartment. For compartment B, taxa 2, 22 and 45 are centrally located, indicating their importance to compartmental interactions, particularly in comparison with those taxa around the periphery of compartment B, such as 21, 33 and 41. Central to compartment A is a food chain (1, 6 and 14). Of the 25 other taxa in A, 23 interact with one of these three taxa. Peripheral taxa placed near to another compartment relate more strongly to that compartment than peripheral taxa placed further away. For example, peripheral taxon 16 has two interactions within compartment A and interacts with taxa that only interact within A, so 16's position is far from B. Conversely, peripheral taxon 41 has one interaction within compartment B and one that goes to A, so its position within B is close to A. Taxon 36 has the role of a bridging taxon between

compartments A and B, where most of the interactions between A and B are attributable to 36.

Because previous work relates compartments to stability¹⁻⁵, we developed a hypothesis of stability by simulating two disturbance scenarios on our food web in Fig. 1. The first removes weakfish (36), which could occur with overfishing. This disturbance¹⁵ translates to a decrease in the number of taxa (-2% = taxa loss). Our overall IC, a variable of interest^{15,17}, was -1.5 -fold the taxa loss. The IC within compartment A, where 36 was a member, was -3 -fold the taxa loss, whereas the IC within B showed no change. Between IC was 18-fold the taxa loss because 36 was a dominant bridging taxon. In our second scenario we considered replacing *Acartia tonsa* (6), a central taxon in compartment A, with an invading zooplankton, a peripheral taxon, that preys only on large bacteria (5) and macrociliates (8) in A and is unpalatable to predators. This new invader reduces realized interactions (-5% = interaction loss). In response to this disturbance, the overall IC changed by the same percentage (1-fold the interaction loss, as expected), as did the IC within compartment A. The IC within compartment B was -0.3 -fold the interaction loss and the between IC was 1.5-fold the interaction loss. Because the factorial change in IC values in relation to the disturbance index (taxa or interaction loss) indicates resistance¹⁵, we propose that compartment B would be the most resistant and the exchange between A and B would be the least resistant to both disturbance events. That is, compartmentalization retains the impacts of a disturbance within a single compartment, minimizing impacts on other compartments and thus providing the stabilizing structure to food webs. This hypothesis is consistent with previous studies¹⁶ that found that weak interactions buffer the effect of disturbances, demonstrated by compartment B's resistance to disturbances occurring in A.

An empirical test²⁹ of hypotheses such as ours requires longitudinal data for multiple food webs, compartmentalized and uncompartimentalized¹⁻⁵, that have undergone disturbance events¹⁵.

Table 2 Compartment analysis for five food webs

Name	n	Weight of interaction	Odds ratio	P	Number of compartments	Overall IC	Within IC	Between IC
Ythan Estuary	134	None	4.19	≥ 0.999	3	0.033	—	—
Little Rock Lake	92	None	3.14	≥ 0.999	2	0.12	—	—
	181	None	10.04	$\leq 0.001^*$	4	0.072	0.17	0.020
St Martin Island	44	None	4.20	≤ 0.907	5	0.11	—	—
	44	Frequency	28.93	≤ 0.803	6	0.0065	—	—
Chesapeake Bay	33	None	8.61	≤ 0.751	3	0.067	—	—
	33	Strength	642.55	$\leq 0.001^*$	2	0.0029	0.0059	0.000093
	33	Carbon	618.75	$\leq 0.012^*$	2	0.0035	0.0071	0.000012
	45	None	9.63	≤ 0.200	4	0.069	—	—
	45	Strength	114.92	$\leq 0.001^*$	2	0.0052	0.0099	0.000087
	45	Carbon	165.84	$\leq 0.001^*$	3	0.0018	0.0044	0.000026
Cypress wetland								
Dry season	64	None	5.52	≤ 0.285	2	0.11	—	—
	64	Strength	11.93	≤ 0.918	3	0.0021	—	—
	64	Carbon	228.18	$\leq 0.001^*$	5	0.00057	0.0020	0.000088
Wet season	64	None	7.81	≤ 0.001	1	0.11	—	—
	64	Strength	12.34	≤ 0.910	4	0.0026	—	—
	64	Carbon	384.81	$\leq 0.001^*$	3	0.00044	0.0013	0.000033

See the text and Table 1 for the calculations of the odds ratio, overall IC and definitions of A, B, C and D. IC within compartments is calculated as $D/(C+D)$, and IC between compartments is calculated as $B/(A+B)$. Asterisk indicates significance at $\alpha = 0.05$.

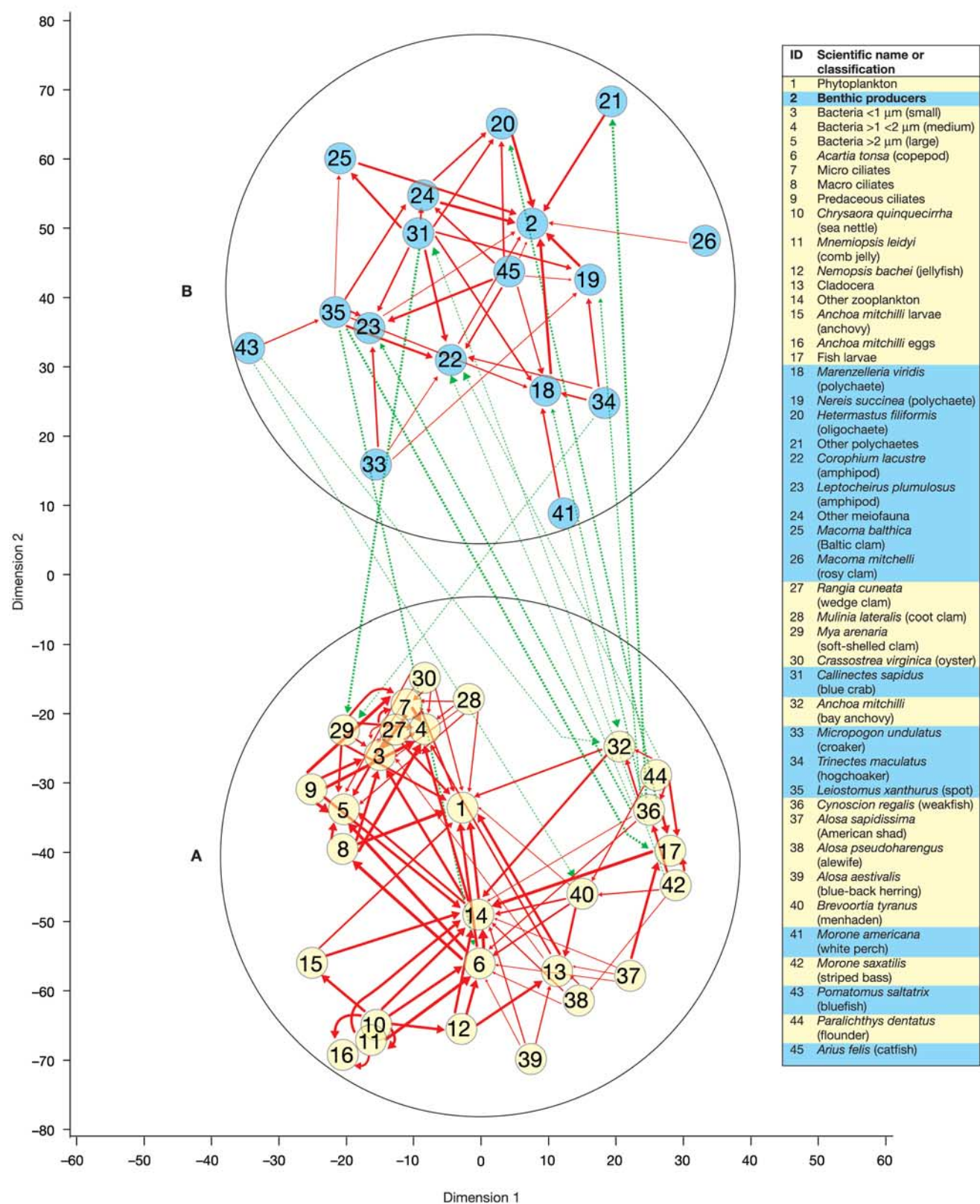


Figure 1 Graphical display of the results for the Chesapeake Bay food web with 45 taxa and weighted by interaction strength. Units are relative distances based on the inverse of the density of interactions. Within-compartment distances were decreased by a factor of 6.2 for aesthetic purposes. Circles indicate compartment boundaries and numbers

identify taxa (yellow, within compartment A; blue, within compartment B). Arrows indicate interactions between taxa (solid red, within compartment; dashed green, between compartments; thickness indicates rank of associated interaction strength) and point from predator to prey.

The changes in variables of interest (for example, IC) should be assessed for stability properties, such as resistance¹⁵. Our results show that this method is a rigorous and effective way to analyse food-web structure and provide the initial steps in understanding the relationship between compartments and stability^{1–5}. □

Methods

Interactions were weighted by interaction frequency, by carbon flow or by interaction strength. Interaction frequency was estimated as acts of predation per hectare per day²⁶. Carbon flow from prey i' to predator i , $W_{ii'}$, was estimated as $\text{g C m}^{-2} \text{ yr}^{-1}$ (refs 27, 28). Interaction strength was the geometric mean of the interaction strengths between predator i and prey i' (ref. 30). The interaction strength of predator i on prey i' was measured as $-(W_{ii'}/B_i)$, where B_i is the biomass of predator i (g C m^{-2}), and the interaction strength of prey i' on predator i was measured as $R_i(W_{ii'}/B_{i'})$, where $B_{i'}$ is the biomass of prey i' (g C m^{-2}) and R_i is the production to consumption ratio of predator i (ref. 30). With weights, the definition of IC becomes the proportion of possible interactions, each with maximum weight, that are realized (with no cannibalistic interactions).

We used the software KliqueFinder¹² to identify compartments. Because KliqueFinder operates on an integer scale from 1 to 99,999 we modified the interactions for weighted versions of the food webs to fit the scale. The modifications had little or no effect on the results. To test whether the concentration of interactions within identified compartments was greater than what was likely to have occurred by chance alone, for each web we conducted Monte Carlo simulations. First, we randomly reassigned interactions, constraining the row marginal (sum of each row in a matrix) to be equal to the row marginal of the original food web, where rows represented predators and columns represented prey. We then applied KliqueFinder and recorded the odds ratio. We then repeated this process 1,000 times to obtain a sampling distribution against which we could compare the empirical odds ratio. Our constraints ensured that the simulated food webs had the same number of predators, the same number (and weight) of interactions associated with a predator, and the same total number (and weight) of realized interactions as the original food web. Basal taxa (taxa with no prey) did vary in simulations for some food web versions, which did change the overall IC from the original for some food webs but we found that there was little or no effect on the P values and no effect on statistical inference.

Although the range of calibrating simulations^{12,14} (the third feature of the methodology) did not allow a direct assessment of the performance of the algorithm for our data (because of large n and weighted interactions), the algorithm typically performs well when there is evidence of compartments in non-weighted data¹². Cannibalism and taxa that interacted with only one other taxon were dropped from the analysis when optimizing the odds ratio because these interactions do not add information to the relative assignments of taxa to compartments. Dropped taxa were added back to the food web for the calculations in Table 2.

Coordinates for the diagram in Fig. 1 were generated by employing multidimensional scaling within and between subgroups¹³, and SAS proc gplot was used to generate the figure. Because of the large magnitude of the difference between the smallest and largest interaction strength weightings ($\sim 10,000$ -fold), the lines were weighted by the rank of the associated interaction strength, where the smallest interaction strength was given a rank of 1 and the largest a rank of 137 (the maximal number of realized interactions). The units are based on the inverse of the between-compartment density (0.0097).

In our scenarios for exploring compartments and stability, we made one simple assumption: the predators of the taxon involved in the disturbance compensated for the loss in their interactions by increasing their interaction strength with their remaining prey items. For the first hypothetical scenario, all interactions with taxon 36 were removed and the predators on 36 had their interaction strengths associated with 36 redistributed proportionally to their other prey interactions. For the second hypothetical scenario, all interactions associated with taxon 6 were removed except for those with taxa 5 and 8. The interactions of the predators on 6 were modified in the same manner as in the first scenario.

Received 12 June; accepted 10 October 2003; doi:10.1038/nature02115.

- Pimm, S. L. The structure of food webs. *Theor. Popul. Biol.* **16**, 144–158 (1979).
- May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, 1973).
- Simon, H. A. in *General Systems: Yearbook of the Society for General Systems* (eds von Bertalanffy, L. & Rapoport, A.) Vol. 10 63–76 (Society for General Systems, Ann Arbor, Michigan, 1965).
- McNaughton, S. J. Stability and diversity of ecological communities. *Nature* **274**, 251–252 (1978).
- Simmel, G. *The Sociology of Georg Simmel* (transl. and ed. Wolff, K. H.) (Free Press, Glencoe, Illinois, 1950).
- Pimm, S. L. & Lawton, J. H. Are food webs divided into compartments? *J. Anim. Ecol.* **49**, 879–898 (1980).
- Yodiz, P. The compartmentation of real and assembled ecosystems. *Am. Nat.* **120**, 551–570 (1982).
- Raffaelli, D. & Hall, S. J. Compartments and predation in an estuarine food web. *J. Anim. Ecol.* **61**, 551–560 (1992).
- Dicks, L. V., Corbet, S. A. & Pywell, R. F. Compartmentalization in plant–insect flower visitor webs. *J. Anim. Ecol.* **71**, 32–43 (2002).
- Moore, J. C. & Hunt, H. W. Resource compartmentation and the stability of real ecosystems. *Nature* **333**, 261–263 (1988).
- Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 8271–8276 (2002).
- Frank, K. Identifying cohesive subgroups. *Soc. Networks* **17**, 27–56 (1995).
- Frank, K. Mapping interactions within and between cohesive subgroups. *Soc. Networks* **18**, 93–119 (1996).
- Frank, K. & Yasumoto, J. Y. Linking action to social structure within a system: social capital within and between groups. *Am. J. Sociol.* **104**, 642–686 (1998).

- Grimm, V. A down-to-earth assessment of stability concepts in ecology: dreams, demands, and the real problems. *Senckenbergiana maritima* **27**, 215–226 (1996).
- McCann, K. S. The diversity–stability debate. *Nature* **405**, 228–233 (2000).
- Dunne, J. A., Williams, R. J. & Martinez, N. D. Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecol. Lett.* **5**, 558–567 (2002).
- Johnson, J. C., Borgatti, S. P., Luczkovich, J. J. & Everett, M. G. Network role analysis in the study of food webs: an application of regular role coloration. *J. Soc. Structure* **2**; published online at (<http://zeeb.library.cmu.edu:7850/JoSS/johnson/RoleAnalysis.html>) (2001).
- McMahon, S. M., Miller, K. H. & Drake, J. Networking tips for social scientists and ecologists. *Science* **293**, 1604–1605 (2001).
- Simmel, G. *Conflict and the Web of Group Affiliations* (transl. Wolff, K. (Free Press, Glencoe, Illinois, 1955).
- Blau, P. M. *Inequality and Heterogeneity* (Macmillan, New York, 1977).
- Martinez, N. D. Artifacts or attributes? Effects of resolution on the Little Rock Lake food web. *Ecol. Monogr.* **61**, 367–392 (1991).
- Raffaelli, D. From Elton to mathematics and back again. *Science* **296**, 1035–1037 (2002).
- Huxham, M., Beaney, S. & Raffaelli, D. Do parasites reduce the chances of triangulation in a real food web? *Oikos* **76**, 284–300 (1996).
- Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).
- Goldwasser, L. & Roughgarden, J. Construction and analysis of a large Caribbean food web. *Ecology* **74**, 1216–1233 (1993).
- Ulanowicz, R. E. & Baird, D. Nutrient controls on ecosystem dynamics: the Chesapeake mesohaline community. *J. Mar. Syst.* **19**, 159–172 (1999).
- Abarca-Arenas, L. G. & Ulanowicz, R. E. The effects of taxonomic aggregation on network analysis. *Ecol. Modell.* **149**, 285–296 (2002).
- Polis, G. A. Complex trophic interactions in deserts: an empirical critique of food-web theory. *Am. Nat.* **138**, 123–155 (1991).
- Neutel, A. M., Heesterbeek, J. A. P. & de Ruiter, P. C. Stability in real food webs: weak links in long loops. *Science* **296**, 1120–1123 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Huxham and D. Raffaelli for providing data on the Ythan Estuary food web; N. Martinez and J. Dunne for providing data on the Little Rock Lake food webs; L. Abarca-Arenas for providing data on the 45-taxa Chesapeake Bay food web; C. Darnell for enhancing the diagram; and C. Goddard and J. Liu for comments and suggestions. This work was supported by the Great Lakes Fishery Commission (A.E.K., D.M.M.), the National Institute of Child Health and Human Development (K.A.E.) and the National Science Foundation (K.A.E.). Opinions reflect those of the authors and do not necessarily reflect those of the granting agency.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.E.K. (krausean@msu.edu).

Activation of the TRPC1 cation channel by metabotropic glutamate receptor mGluR1

Sang Jeong Kim^{1,2}, Yu Shin Kim¹, Joseph P. Yuan¹, Ronald S. Petralia³, Paul F. Worley¹ & David J. Linden¹

¹Department of Neuroscience, Johns Hopkins University School of Medicine, 725 N. Wolfe Street, Baltimore, Maryland 21205, USA

²Department of Physiology, Kangwon National University School of Medicine, Chuncheon, 200-71, Korea

³NIDCD/NIH, Building 50, Room 4142 50 South Drive, MSC 8027, Bethesda, Maryland 20892, USA

Group I metabotropic glutamate receptors (consisting of mGluR1 and mGluR5) are G-protein-coupled neurotransmitter receptors¹ that are found in the perisynaptic region of the post-synaptic membrane². These receptors are not activated by single synaptic volleys but rather require bursts of activity^{3–5}. They are implicated in many forms of neural plasticity including hippocampal long-term potentiation and depression^{6–8}, cerebellar long-term depression^{8–11}, associative learning^{7,11}, and cocaine addiction¹². When activated, group I mGluRs engage two G-protein-dependent signalling mechanisms: stimulation of phospholipase C and activation of an unidentified, mixed-cation

excitatory postsynaptic conductance (EPSC), displaying slow activation, in the plasma membrane^{4,5,13–15}. Here we report that the mGluR1-evoked slow EPSC is mediated by the TRPC1 cation channel. TRPC1 is expressed in perisynaptic regions of the cerebellar parallel fibre–Purkinje cell synapse and is physically associated with mGluR1. Manipulations that interfere with TRPC1 block the mGluR1-evoked slow EPSC in Purkinje cells; however, fast transmission mediated by AMPA-type glutamate receptors remains unaffected. Furthermore, co-expression of mGluR1 and TRPC1 in a heterologous system reconstituted a mGluR1-evoked conductance that closely resembles the slow EPSC in Purkinje cells.

In cerebellar Purkinje cells, mGluR1 is linked via a $G\alpha_q$ protein complex to the activation of phospholipase C β (PLC- β), which cleaves the membrane phospholipid phosphatidylinositol-4,5-bisphosphate to yield 1,2-diaclyglycerol and inositol-1,4,5-trisphosphate (InsP₃). Activation of mGluR1 by burst stimulation of parallel fibres results in both Ca²⁺ mobilization through a PLC- β –InsP₃ cascade^{16–18} and activation of a slow EPSC³ carried by a mixed-cation conductance^{4,5,13–15}. The identity of the ion channel underlying the mGluR1-evoked slow EPSC has not been determined in Purkinje cells or in any other cell type. There are, however, a number

of observations that constrain the identity of the ion channel underlying the slow EPSC. It is a mixed-cation conductance that reverses at about +20 mV (refs 5, 14, 15) and requires external Ca²⁺ (refs 5, 19). It is not blocked by antagonists of Na⁺/Ca²⁺ exchangers, purinergic receptors, hyperpolarization-activated cation channels or voltage-gated Ca²⁺ channels^{5,13,15}.

There are several reasons to believe that the Ca²⁺ mobilization and slow EPSC signals evoked by mGluR1 activation are dependent on $G\alpha_q$ but subsequently involve divergent pathways. First, drugs that interfere with PLC- β , InsP₃ receptors or internal Ca²⁺ stores completely block Ca²⁺ mobilization but have variable effects on the slow EPSC conductance, ranging from no blockade to weak blockade^{4,5,13–15}. Second, photolysis of caged InsP₃ or Ca²⁺ does not mimic the slow EPSC¹⁵. Third, drugs^{4,13} and induced mutations²⁰ that inhibit $G\alpha_q$ block both Ca²⁺ mobilization and the slow EPSC. Fourth, weak burst stimulation of parallel fibres can evoke Ca²⁺ mobilization in the absence of the slow EPSC, whereas stronger burst stimulation recruits both signals (ref. 16 and S.J.K., P.F.W. and D.J.L., unpublished observations).

Group I mGluRs are organized at glutamatergic synapses through interactions with scaffolding molecules, most notably the Homer family of proteins, which form multimers capable of regulating

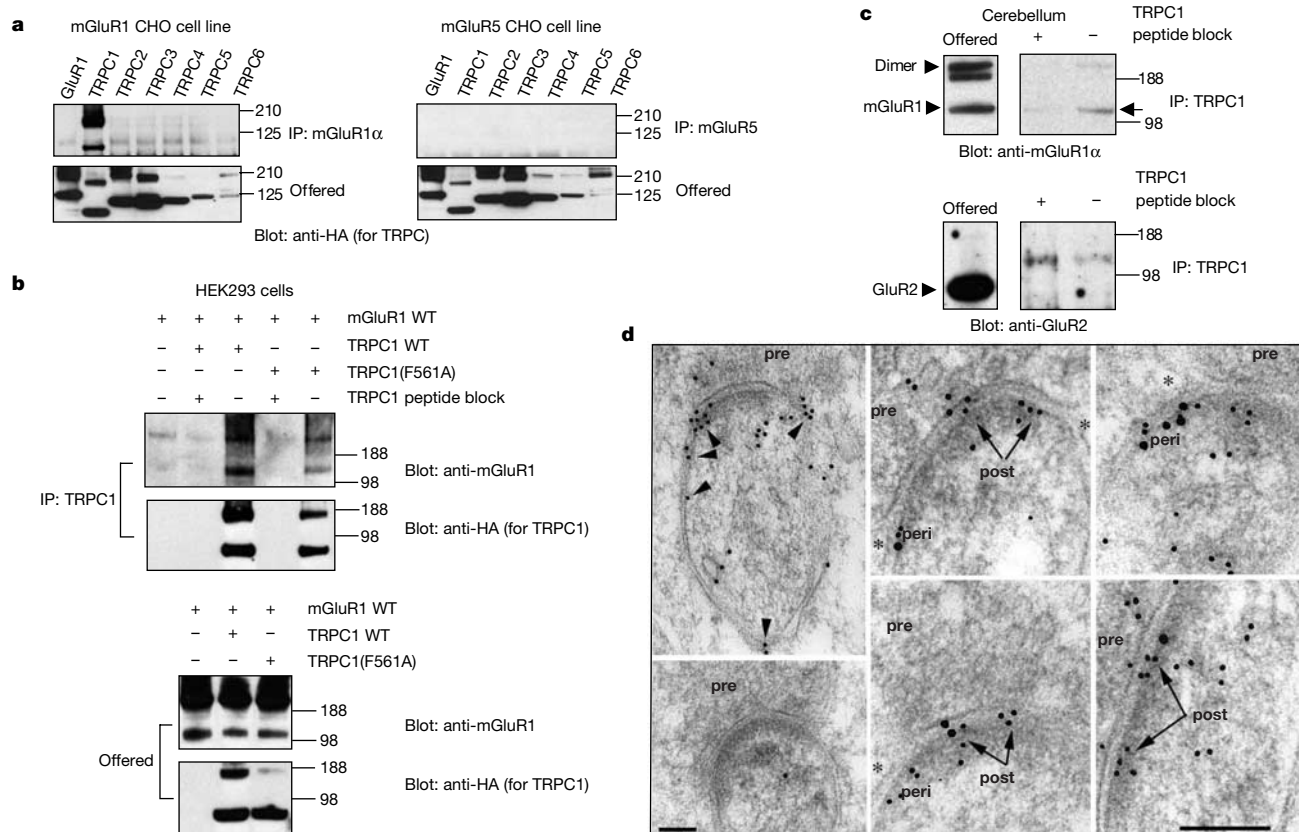


Figure 1 mGluR1 α and TRPC1 physically associate in CHO cells, HEK293 cells and in brain, and they co-localize on Purkinje cell spines. **a**, mGluR1 α - and mGluR5-expressing CHO cell lines were transfected with haemagglutinin (HA)-tagged TRPC1–6. HA–TRPC1, but not GluR1 or other TRPCs, co-immunoprecipitate with mGluR1 α . None of the TRPCs co-immunoprecipitate with mGluR5. Markers are in kDa. **b**, HEK293 cells were transfected with mGluR1 α and/or HA–TRPC1 (WT or F561A). TRPC1 and TRPC1(F561A) co-immunoprecipitate mGluR1 α . Anti-TRPC1 antibody does not immunoprecipitate mGluR1 α in the absence of TRPC1. **c**, TRPC1 co-immunoprecipitates mGluR1 α (arrow) but not GluR2 (overexposed and displaying a nonspecific band greater than the size of GluR2) from cerebellum. Preabsorption of TRPC1 antibody with peptide antigen reduced mGluR1 α by 90%. **d**, The left column shows immunogold labelling for TRPC1 associated

with spine synapses in the molecular layer of the cerebellum. Arrowheads indicate labelling with 10 nm F(ab')₂-gold associated with the perisynaptic as well as extrasynaptic and outer synaptic membrane areas of the postsynaptic spine. Sections were labelled with antibody without (left panel, top row) and with (left panel, bottom row) preabsorption with peptide. Pre, presynaptic terminal. Scale bar: 100 nm. The middle and right columns show co-localization of immunogold labelling for TRPC1 (5 nm) and mGluR1 α (10 nm) associated with spine synapses in the molecular layer of the cerebellum. Arrows indicate postsynaptic labelling (post). Peri refers to perisynaptic/extrasynaptic labelling of the spine membrane near glia processes (asterisks) adjacent to the synaptic cleft. Some labelling for TRPC1 is also seen in the presynaptic terminals (pre). Scale bar: 100 nm.

coupling of group I mGluRs to a number of targets including InsP₃ receptors, ryanodine receptor and voltage-gated calcium channels²¹. In recent studies we found that Homer also binds and regulates the gating of members of the TRPC family of nonspecific cation channels²². TRPC channels are activated in response to G-protein-coupled receptor activation and/or depletion of intracellular Ca²⁺ stores, and are believed to participate in replenishment and regulation of intracellular Ca²⁺ pools in a process termed capacitive calcium entry²³. TRPC1 is expressed in cerebellar Purkinje cells²⁴. Accordingly, we examined the hypothesis that a TRPC ion channel underlies the mGluR1-evoked slow EPSC.

CHO cell lines that stably express mGluR1 α or mGluR5 were transfected with TRPC1, TRPC2, TRPC3, TRPC4, TRPC5 or TRPC6. mGluR1 α antibody co-immunoprecipitated TRPC1 but not other TRPCs, whereas mGluR5 antibody did not co-immunoprecipitate any of the TRPC channels (Fig. 1a). mGluR1 α and TRPC1 also co-immunoprecipitated using a TRPC1 antibody (Fig. 1b). These data indicate a specific physical interaction between TRPC1 and mGluR1 α . A TRPC1 mutant lacking channel activity (described below) showed similar association with mGluR1 α . To determine whether mGluR1 α and TRPC1 associate *in vivo*, we assayed for co-immunoprecipitation of mGluR1 α from detergent lysates of cerebellum using TRPC1 antibodies (Fig. 1c). TRPC1 antibody co-immunoprecipitated mGluR1 α and this was blocked by TRPC1 peptide. TRPC1 antibody did not co-immunoprecipitate

the AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptor subunit GluR2. To assess the ultrastructural localization of TRPC1 we performed post-embedding immuno-electron microscopy on adult rat cerebellum (Fig. 1d). Gold particles preferentially localized to dendritic spines of Purkinje neurons where they co-localized with mGluR1 α along the spine membrane. Co-localization was evident in the perisynaptic region as well as at the postsynaptic membrane. These biochemical and anatomical data make TRPC1 an attractive candidate for the ion channel underlying the mGluR1-evoked slow EPSC. Indeed, this suggestion has been made previously based on the similarity of a current recorded in a heterologous cell expressing TRPC1 and either TRPC4 or TRPC5 (ref. 24) to that recorded in hippocampal²⁵ or midbrain dopaminergic²⁶ neurons stimulated with a group I mGluR agonist.

We next examined the mGluR1-evoked EPSC in Purkinje neurons of acute cerebellar slices using whole-cell patch clamp. Stimulation of parallel fibres with a brief burst gives rise to a fast component predominantly mediated by AMPA receptors and a slow component mediated by mGluR1. In some cases the falling phase of the fast component overlaps the rising phase of the slow component, making it difficult to measure them separately. To facilitate our analysis of the delayed EPSC, we applied the AMPA/kainate receptor antagonist CNQX (6-cyano-7-nitroquinoxaline-2,3-dione) at a concentration that produces an approximately 90% reduction in the amplitude of the fast EPSC (10 μ M). This allowed for clearly

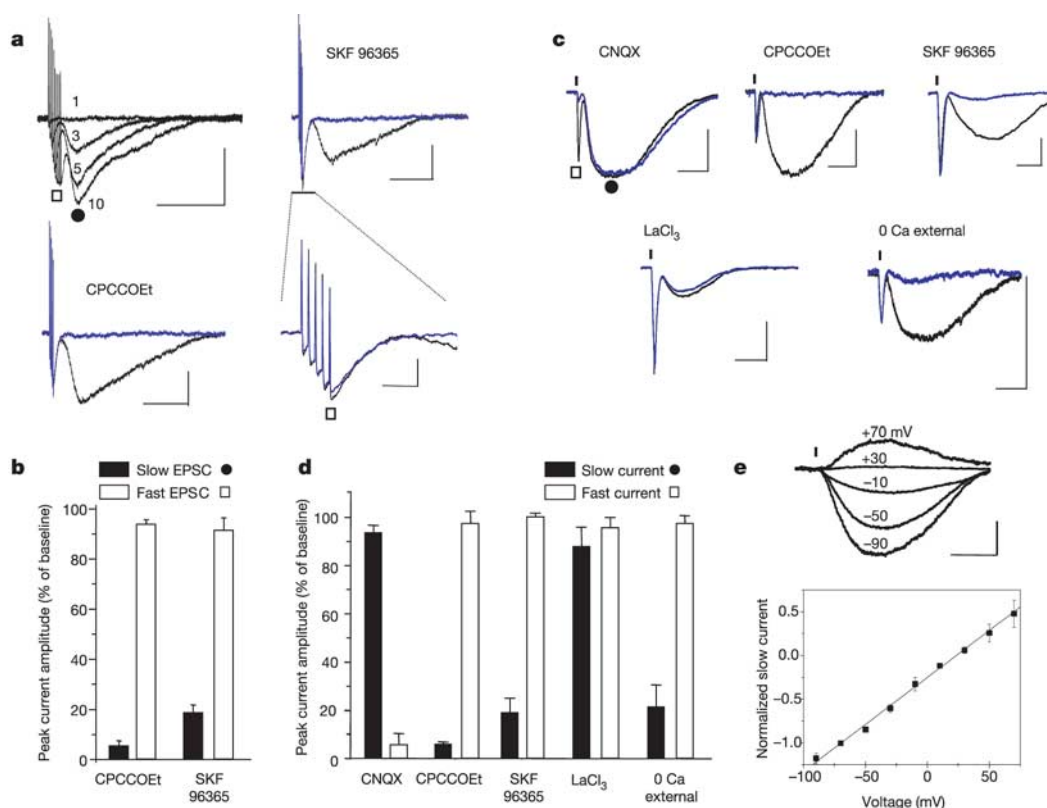


Figure 2 Characterization of the mGluR1-evoked slow EPSC in Purkinje cells.

a, Representative single traces show fast and slow components of the parallel-fibre-evoked EPSC which were produced by burst stimulation at 100 Hz in the presence of CNQX (10 μ M) and GABAzine (5 μ M). A burst consisting of five pulses was adopted as the standard stimulus. The black traces represent the pretreatment (baseline) condition whereas the blue traces represent the post-treatment condition at an asymptotic time point (typically 5–12 min after introduction of the treatment). Numbers in the top left panel indicate number of pulses. Scale bars: 100 pA, 0.5 s, except for the time-expanded traces in the lower right panel (50 ms). **b**, Population data for the manipulations shown in **a** are expressed as the percentage of baseline peak current amplitude and are presented as

mean $\pm$ s.e.m. **c**, Fast and slow inward currents were evoked by micropressure pulses (10 p.s.i., 50 ms duration) of external solution containing DHPG (100 μ M) and glutamate (40–100 μ M) delivered through a patch pipette. Representative current traces are shown before (black) and after (blue) various bath-applied treatments. Scale bars: 200 pA, 0.5 s. **d**, Population data for the manipulations shown in **c** are expressed as the percentage of baseline peak current amplitude. **e**, Upper panel shows the DHPG-evoked slow currents at different potentials in the presence of TTX (1 μ M), NBQX (20 μ M), CdCl₂ (200 μ M) and BaCl₂ (5 mM, by replacing CaCl₂). Scale bars: 200 pA, 0.5 s. Lower panel shows current–voltage relations for a population of five Purkinje cells. Current has been normalized to the value at –70 mV. Linear regression line is shown in the plot.

resolvable fast and slow EPSC components (Fig. 2a, b). The amplitude of the stimulation current was adjusted to evoke a 300–500 pA fast EPSC in the absence of CNQX. A burst frequency of 100 Hz was used to maximize the slow EPSC⁶ and the holding potential was set to -70 mV. As previously reported^{3,4}, single stimuli of parallel fibres did not evoke a measurable slow EPSC, and the amplitude of the slow EPSC depended on the number of pulses in the train. Bath application of an mGluR1 antagonist, CPCCOEt (100 μ M), completely abolished the slow EPSC evoked by a train of five stimuli ($4.9 \pm 1.9\%$ of baseline, mean $\pm$ s.e.m., $n = 7$ cells, within-cell comparisons). A nonselective antagonist of receptor-operated cation channels, SKF 96365 (30 μ M), also strongly blocked the mGluR-evoked slow EPSC ($18.6 \pm 3.0\%$ of baseline, $n = 13$ cells). The inhibitory effects of both CPCCOEt and SKF 96365 on the slow EPSC were selective, having no effect on the peak amplitude of the fast EPSC ($93.8 \pm 1.8\%$, $n = 7$ cells; $90.2 \pm 6.4\%$, $n = 10$ cells, respectively).

To evaluate specifically postsynaptic responses, and to avoid possible effects of manipulations on presynaptic release, mGluRs were activated by micropressure pulses of 3,5-dihydroxyphenylglycine (DHPG, 100 μ M; 10 pounds per square inch (p.s.i.), 50 ms) delivered through a patch pipette. Glutamate (40–100 μ M) was included in this solution to produce a small, fast inward current, serving as a control measure for nonspecific effects (Fig. 2c, d). Similar to results with the EPSC evoked by burst stimulation of parallel fibres, CPCCOEt (100 μ M) and SKF 96365 (30 μ M) blocked

the mGluR-evoked slow current ($5.9 \pm 1.0\%$, $n = 6$ cells; $18.9 \pm 6.1\%$, $n = 8$ cells, respectively) but left the fast current unaffected. LaCl₃, which blocks voltage-gated calcium channels and augments some, but not all, TRPC channels in heterologous expression systems²⁴, did not significantly attenuate either the fast or slow component of the inward current ($95.7 \pm 4.3\%$; $88.0 \pm 8.0\%$, $n = 5$ cells, respectively). Removal of external calcium attenuated the slow current ($21.5 \pm 9.0\%$, $n = 5$ cells) but did not affect the fast current ($97.5 \pm 6.4\%$, $n = 5$ cells). The current–voltage relation of the slow current in Purkinje cells was roughly linear and reversed at about $+23$ mV, consistent with a previous report¹⁵. These properties of the mGluR1-evoked slow current in Purkinje cells are consistent with a conductance mediated by TRPC1, but they do not uniquely implicate TRPC1, nor do they exclude some other candidate ion channels.

To examine the role of TRPC1 in the mGluR1-evoked current, Purkinje neurons in slice culture were transfected with an expression construct of the channel mutant TRPC1(F561A) (Fig. 1b) along with enhanced green fluorescent protein (EGFP; Fig. 3a). In control cells transfected with EGFP alone and stimulated with micropressure pulses of DHPG plus glutamate, the peak amplitude of the slow current was 640 ± 72 pA ($n = 10$ cells). In cells expressing TRPC1(F561A), the amplitude of mGluR-evoked slow current was reduced to 49% of that measured in an EGFP-alone control (313 ± 39 pA, $n = 33$ cells, $P < 0.001$). The amplitude of slow currents in non-transfected Purkinje cells (686 ± 95 pA, $n = 9$ cells) and in cells transfected with wild-type TRPC1 (802 ± 116 pA, $n = 11$ cells) was significantly larger than that of cells transfected with TRPC1(F561A) ($P < 0.001$ for both), and was not significantly different from cells transfected with EGFP alone ($P > 0.10$ for both).

To provide an independent test for the role of TRPC1, we included TRPC1 antibody (30 μ g ml⁻¹) in the pipette solution and performed whole-cell recordings from Purkinje cells in organotypic culture (Fig. 3b). In a separate group of cells we applied the antibody rendered nonfunctional by preabsorption with TRPC1 peptide. In TRPC1-antibody-treated cells the amplitude of the slow current started to decrease about 15 min after achieving stable whole-cell recording. At 37.5 min, the difference between the normalized slow current with antibody preabsorbed with TRPC1 peptide (nonfunctional; 1.02 ± 0.06 , $n = 6$ cells) and that with TRPC1 antibody (0.36 ± 0.06 , $n = 6$ cells) was statistically significant ($P < 0.01$). This time course probably reflects the diffusion of TRPC1 antibody from the pipette to the dendritic site of mGluR1 activation.

As a further test of the hypothesis that TRPC1 underlies the mGluR1-evoked slow EPSC, we attempted to reconstitute this conductance in a heterologous system. A line of CHO cells constitutively expressing mGluR1 was co-transfected with EGFP and TRPC1. EGFP-positive cells were subjected to whole-cell patch-clamp recording. A command potential of -70 mV was imposed, and mGluR1 was activated with a micropressure pulse of DHPG (10 p.s.i., 100 ms). This evoked a slowly activating inward current with a peak current density of 19.3 ± 3.9 pA pF⁻¹ (mean $\pm$ s.e.m., $n = 8$ cells; Fig. 4a, b). TRPC5 is reported to form heteromultimers with TRPC1 (ref. 27), and CHO cells transfected with both TRPC1 and TRPC5 displayed an even larger DHPG-evoked current (50.7 ± 7.8 pA pF⁻¹, $n = 7$ cells). In contrast, when a separate population of CHO cells was transfected with TRPC3—a related channel that does not form a complex with mGluR1—no significant inward current was evoked (0.5 ± 0.5 pA pF⁻¹ and 0.2 ± 0.3 pA pF⁻¹, respectively, $n = 6$ cells per group). Transfection with TRPC1(F561A) similarly yielded no significant current (0.4 ± 0.3 pA pF⁻¹, $n = 7$). When TRPC1 was transfected together with TRPC1(F561A) in a ratio of 2:1, the amplitude of the evoked slow current was reduced to about 25% of that seen with TRPC1 alone (4.8 ± 2.5 pA pF⁻¹, $n = 6$). This is similar to the effect of

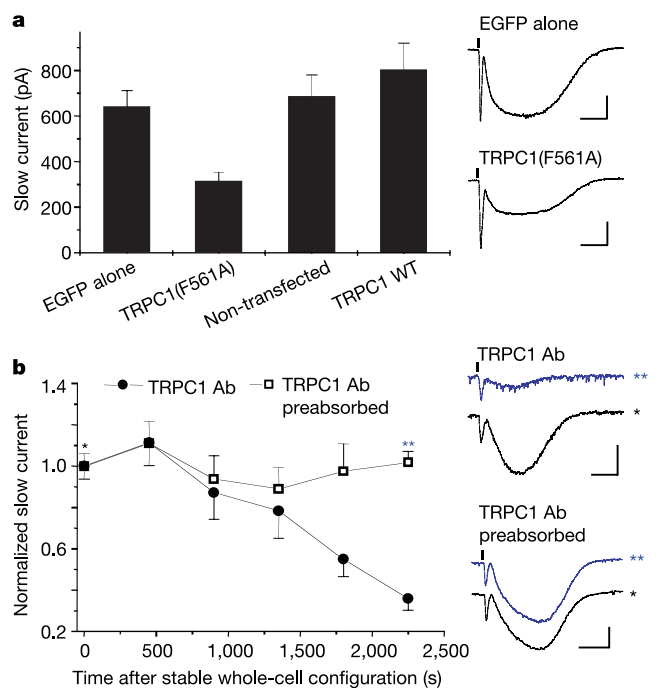


Figure 3 Manipulations that interfere with TRPC1 function block the mGluR1-evoked slow current in Purkinje cells. **a**, Population data in four different conditions are expressed as the peak amplitude of mGluR1-evoked slow current in Purkinje cells in organotypic slice cultures. The amplitude of slow current from cells transfected with a dominant-negative pore mutant of TRPC1 (TRPC1(F561A)) was attenuated in comparison to that of cells transfected with EGFP alone, non-transfected cells and cells transfected with wild-type TRPC1. Representative single-current traces are shown to the right. Scale bars: 200 pA, 0.5 s. **b**, Time course of the amplitude of the slow current that was activated by micropressure pulses of DHPG together with glutamate. The current amplitude was normalized to the value recorded immediately after achieving a stable whole-cell configuration. Representative current traces are shown to the right. Scale bars: 200 pA, 0.5 s. Ab, antibody. Representative current traces are shown before (black) and after (blue) various treatments.

TRPC1(F561A) transfection in Purkinje cells (Fig. 3). The current-voltage relation for TRPC1/TRPC5-transfected CHO cells (not shown) revealed that the DHPG-evoked current was roughly linear and reversed at about +23 mV, similar to the mGluR1 EPSC in Purkinje cells (Fig. 2 and ref. 22). Both CPCCOEt (100 μ M) and SKF 96365 (50 μ M) completely abolished the DHPG-evoked current ($5 \pm 4\%$ of baseline, $n = 7$ cells; $3 \pm 3\%$ of baseline, $n = 8$ cells, respectively). The DHPG-evoked current was not significantly altered by LaCl_3 (100 μ M; $91 \pm 9\%$ of baseline, $n = 7$ cells). Removal of external calcium, previously shown to attenuate the slow EPSC in Purkinje cells^{7,28} (see Fig. 2c, d), produced a near-complete blockade of the DHPG-evoked current ($13 \pm 4\%$ of baseline, $n = 7$ cells). Notably, internal application of a TRPC1-inactivating antibody (30 $\mu\text{g ml}^{-1}$), but not a control solution consisting of antibody preabsorbed with TRPC1 peptide, produced a strong attenuation of DHPG-evoked current ($26 \pm 4\%$ of baseline, $n = 6$; $109 \pm 8\%$ of baseline, $n = 7$, respectively).

Biochemical studies demonstrate a selective interaction of mGluR1 with TRPC1 (Fig. 1a). To assess whether mGluR1 may be selectively capable of activating TRPC1, we performed simultaneous voltage-clamp and ratiometric calcium imaging of CHO cells expressing TRPC1 with other G-protein receptors, including mGluR5 or M1 muscarinic receptor. In cells expressing mGluR5 and TRPC1, DHPG application resulted in a robust calcium transient (568 ± 155 nM, peak free calcium, $n = 10$ cells; thereby

confirming that mGluR5 was functionally expressed) but no significant activation of membrane current (0.9 ± 0.8 pA pF⁻¹). When this experiment was repeated with cells expressing mGluR1, strong calcium responses (783 ± 145 nM, $n = 10$) and a robust inward current (25.3 ± 4.6 pA pF⁻¹) were observed, similar to that previously seen using a CHO cell line engineered to stably express mGluR1 (Supplementary Fig. 1). When this experiment was repeated with cells expressing M1 muscarinic receptor in place of mGluR, and stimulation with the muscarinic agonist pilocarpine in place of DHPG, the results were similar to that seen with mGluR5 (that is, a strong calcium signal was observed (696 ± 88 nM, $n = 10$) but no significant membrane current accompanied this calcium transient (0.6 ± 0.5 pA pF⁻¹)). These findings show that not all G_q/PLC-coupled receptors can activate the TRPC1-mediated slow conductance, and they suggest that the specific physical association of mGluR1 and TRPC1 is important for signal coupling.

Our observations indicate that TRPC1 is required for the mGluR1-induced inward current in Purkinje neurons. Because TRPC1 is known to co-assemble with TRPC4 and TRPC5, it is possible that the native channel is a heteromultimer^{24,27}. Furthermore, although it is likely that the slow EPSC evoked by group I mGluR activation in Purkinje cells is similar to that in other neurons, there may be some important differences as well. For example, the slow EPSC recorded in hippocampal pyramidal neurons seems to require activation of both mGluR1 and

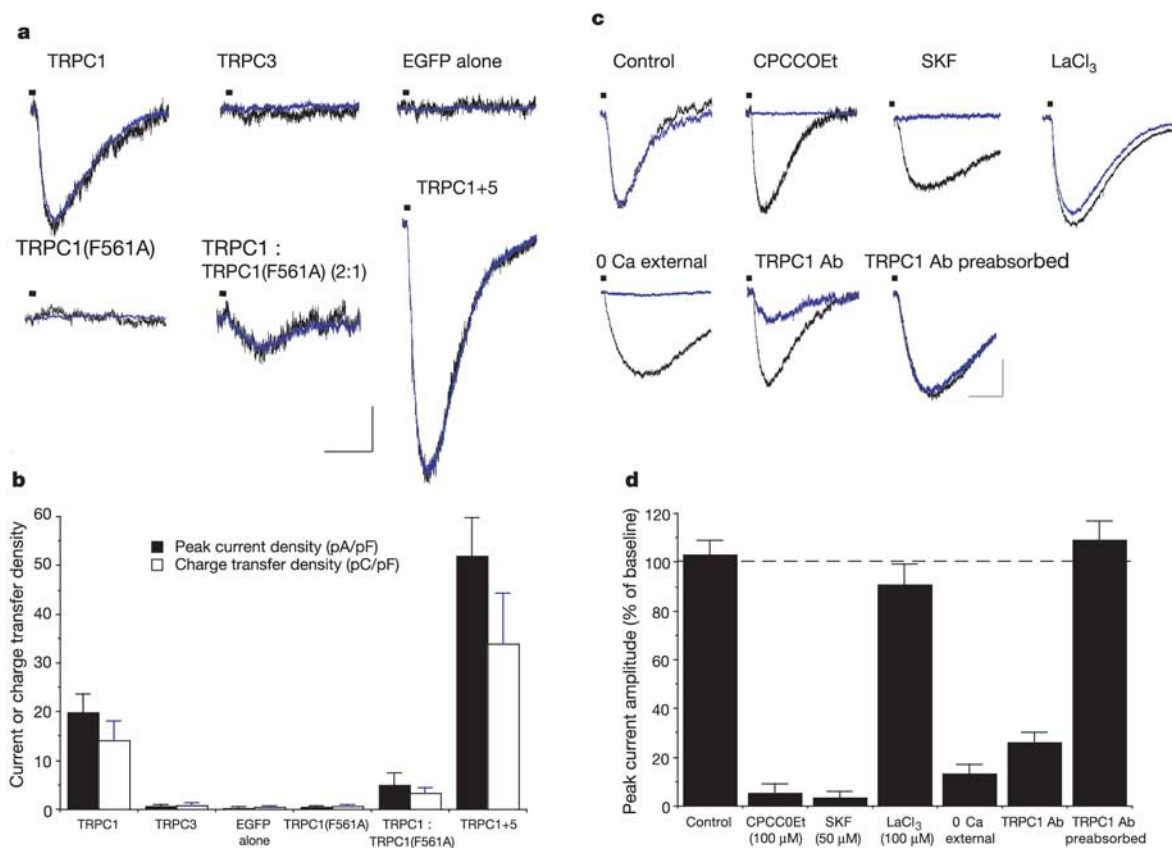


Figure 4 The inward current evoked by DHPG in mGluR1/TRPC1-expressing CHO cells closely resembles the mGluR1-evoked slow EPSC in Purkinje cells. **a**, Representative current traces in a line of CHO cells that constitutively expressed mGluR1 together with the indicated ion channels. Micropressure pulses of DHPG were delivered as indicated by the horizontal bars (100 ms, 10 p.s.i.). Black traces are single sweeps and blue traces are the average of six consecutive sweeps. Scale bars: 50 pA, 1 s. **b**, Population data for the manipulations shown in **a** normalized to cell membrane capacitance (as an index of plasma membrane surface area) and presented as mean $\pm$ s.e.m. **c**, Representative

current traces for experiments using within-cell comparisons for CHO cells expressing TRPC1 together with mGluR1. In this panel all traces are the average of six consecutive sweeps. The black trace represents the pre-treatment condition whereas the blue trace represents the post-treatment condition at an asymptotic time point (typically 5–12 min after introduction of the treatment). Scale bars: 50 pA, 1 s. **d**, Population data for the manipulations shown in **c** are expressed as the percentage of baseline peak current amplitude.

mGluR5 to produce a maximal effect²⁵ whereas Purkinje cells only express mGluR1.

What is the function of the mGluR-evoked slow EPSC? One possibility is that the Ca^{2+} influx associated with this current serves to replenish internal stores depleted by activation of the PLC–InsP₃ signalling cascade. In other cell types this is accomplished by direct coupling between calcium stores and TRPC channels²³. In Purkinje cells the relationship between activation of the PLC–InsP₃ cascade and the TRPC1 cascade is less clear, as manipulations that interfere with PLC- β , InsP₃ receptors or internal Ca^{2+} stores have weak and variable effects on the slow EPSC conductance^{4,5,13–15}. It is also possible that the Ca^{2+} and/or Na^{+} influx mediated by the slow EPSC engages other second messenger pathways. For example, activation of mGluR1 in Purkinje cells drives the postsynaptic production of endocannabinoids, which function as diffusible retrograde messengers to transiently depress transmitter release²⁸. □

Methods

Cell biology assays

Detergent (1% Triton) lysates of rat cerebellum and cell lines were generated and analysed as described²². TRPC constructs, cell culture and transfection methods are as described²². A total of 100 μl lysates precipitated with 2 μg of anti-TRPC1 antibodies (Alomone). For the negative control, 2 μg of anti-TRPC1 antibodies were preabsorbed with equal amounts of TRPC1 peptide antigen (amino acids 523–537 TRPC1) for 6 h at 4 °C. TRPC1 point mutant F561A was generated by QuikChange site-directed mutagenesis (Stratagene).

Immuno-electron microscopy and immunogold quantification

Immuno-electron microscopy was performed by post-fixation immunogold labelling²⁹ using a TRPC1 antibody (Alomone) at a dilution of 1:100. A monoclonal mGluR1 α antibody²⁹ was used for double labelling. Brightness and contrast of micrographs were modified using Adobe Photoshop. Results of immunogold labelling are shown in Fig. 1d. Pretreatment of the TRPC1 antibody (left panel of bottom row) with the antigen peptide removed 95% of the gold labelling. Counts of gold particles were 2.67 gold particles per synapse for the area of the postsynaptic density (PSD) and 2.80 for the remaining extrasynaptic area of the spine (63 spine synapses from two animals). As the PSD represents about one-third of the area of spine profiles, these counts indicate enrichment of labelling near the PSD. In co-localization studies a total of 96% of spines labelled with mGluR1 also displayed label for TRPC1 on the spine membrane (Fig. 1d, $n = 121$; two animals). The postsynaptic membrane contained 51% and 35% of TRPC1 and mGluR1 α membrane labelling, respectively. The perisynaptic region (defined as one-third of the length of the postsynaptic membrane) contained 17% and 38% of TRPC1 and mGluR1 α membrane labelling, respectively.

Electrophysiology

Parasagittal slices of the cerebellar vermis (200 μm thick) were prepared from P18–20 Sprague–Dawley rats using a vibrating tissue slicer and ice-cold standard artificial cerebrospinal fluid (ACSF) containing 124 mM NaCl, 5 mM KCl, 1.25 mM Na_2HPO_4 , 2 mM MgSO_4 , 2 mM CaCl_2 , 26 mM NaHCO_3 and 10 mM D-glucose, bubbled with 95% O_2 and 5% CO_2 . After a recovery period of 30 min at 37 °C, the slices were placed in a submerged chamber that was perfused at a rate of 2 ml min⁻¹ with ACSF supplemented with 5 μM GABAzine to block GABA_A receptors. Experiments were performed using the visualized whole-cell patch-clamp technique at room temperature. The recording electrodes (resistance 4–5 M Ω) were filled with a solution containing 135 mM Cs-methanesulphonate, 10 mM CsCl, 10 mM HEPES, 4 mM Na_2ATP , 0.4 mM Na_3GTP and 0.2 mM EGTA (pH 7.25). Calcium-free extracellular solution was prepared by replacing CaCl_2 with MgCl_2 and adding 0.2 mM EGTA. LaCl_3 was added to a HEPES-based external solution to prevent precipitation. All drugs were purchased from Sigma except for CPCCOEt, SKF 96365, GABAzine and CNQX, which were purchased from Tocris Cookson. Currents were filtered at 1 kHz and digitized at 5 kHz. For parallel fibre stimulation, standard patch pipettes were used and were filled with external saline and placed in the middle third of the molecular layer. Synaptic responses were evoked every 30 s using 12–16 μA pulses (100 μs duration). When burst stimulation was used, the interpulse interval was 10 ms. In some experiments membrane currents were evoked by agonist application using a pressure application system (Picospritzer; 10 p.s.i., 50 ms pulse duration).

For organotypic cerebellar cultures, 200- μm thick parasagittal slices from P10–12 rats were placed in 1 μm Millicell inserts (Becton Dickinson) within sterile 6-well plates. Medium (1.5 ml) was placed under each insert (50% Eagle's basal medium, 20% Hank's balanced salt solution, 25% heat-inactivated horse serum, 30 mM glucose, 5 mM glutamine, 5 mM HEPES, 1 μg ml⁻¹ insulin, 0.012% ascorbic acid, and 100 U ml⁻¹ penicillin-streptomycin; all from GIBCO BRL) to achieve an interface configuration. Plates were incubated in 5% CO_2 in air at 37 °C. Cultured slices were transfected using the Helios Gene Gun (Bio-Rad) 24 h after slice preparation. To prepare 'bullets', 25 mg of 1 μm gold particles (Bio-Rad) were coated with DNA plasmid combinations (45 μg TRPC1 mutant plus 5 μg EGFP, 45 μg TRPC1 wild type plus 5 μg EGFP, and 45 μg PRK5 (empty plasmid) DNA plus 5 μg EGFP DNA). A minimum of 10 μl each of 0.5 M spermidine and CaCl_2 was used, or this volume was increased to match the volume of DNA. Coated, rinsed gold particles were resuspended in 0.06 mg ml⁻¹ polyvinylpyrrolidone in ethanol and then loaded into tubing. Particles were accelerated into cultured slices with the gene gun

pressurized to 180 p.s.i. Plates were returned to the incubator immediately and used for recording after 24 h.

CHO cells were bathed in a solution that contained NaCl (140 mM), KCl (5 mM), CaCl_2 (2 mM), MgCl_2 (1 mM), HEPES (10 mM) and glucose (10 mM), adjusted to pH 7.35 with NaOH, which flowed at a rate of 1 ml min⁻¹. The internal saline was identical to that used in Purkinje cell recordings. Patch electrodes were pulled from N51A glass and yielded a resistance of 3–5 M Ω . Membrane currents were recorded in resistive voltage-clamp mode, filtered at 2 kHz and digitized at 5 kHz. For CHO cell experiments in which voltage-clamp recording and calcium imaging were combined, EGTA was omitted from the internal saline and replaced with 0.2 mM bis-fura-2. Imaging of free calcium was performed as previously described³⁰. Experiments were conducted at room temperature. Cells in which R_{input} or R_{series} varied by more than 15% were excluded from the analysis.

Received 21 February; accepted 30 October 2003; doi:10.1038/nature02162.

Published online 12 November 2003.

- Conn, P. J. & Pin, J.-P. Pharmacology and functions of metabotropic glutamate receptors. *Annu. Rev. Pharmacol. Toxicol.* **37**, 205–237 (1997).
- Nusser, Z., Mulvihill, E., Streit, P. & Somogyi, P. Subsynaptic segregation of metabotropic and ionotropic glutamate receptors as revealed by immunogold localization. *Neuroscience* **61**, 421–427 (1994).
- Batchelor, A. M., Madge, D. J. & Garthwaite, J. Synaptic activation of metabotropic glutamate receptors in the parallel fibre-Purkinje cell pathway in rat cerebellar slices. *Neuroscience* **63**, 911–915 (1994).
- Tempia, F., Miniaci, M. C., Anchisi, D. & Strata, P. Postsynaptic current mediated by metabotropic glutamate receptors in cerebellar Purkinje cells. *J. Neurophysiol.* **80**, 520–528 (1998).
- Tempia, F., Alojado, M. E., Strata, P. & Knopfel, T. Characterization of the mGluR1-mediated electrical and calcium signaling in Purkinje cells of mouse cerebellar slices. *J. Neurophysiol.* **86**, 1389–1397 (2001).
- Bortolotto, Z. A., Fitzjohn, S. M. & Collingridge, G. L. Roles of metabotropic glutamate receptors in LTP and LTD in the hippocampus. *Curr. Opin. Neurobiol.* **9**, 299–304 (1998).
- Aiba, A. *et al.* Reduced hippocampal long-term potentiation and context-specific deficit in associative learning in mGluR1 mutant mice. *Cell* **79**, 365–375 (1994).
- Conquet, F. *et al.* Motor deficit and impairment of synaptic plasticity in mice lacking mGluR1. *Nature* **372**, 237–243 (1994).
- Kano, M. & Kato, M. Quisqualate receptors are specifically involved in cerebellar synaptic plasticity. *Nature* **325**, 276–279 (1987).
- Linden, D. J., Dickinson, M. H., Smeyne, M. & Connor, J. A. A long-term depression of AMPA currents in cultured cerebellar Purkinje neurons. *Neuron* **7**, 81–89 (1991).
- Aiba, A. *et al.* Deficient cerebellar long-term depression and impaired motor learning in mGluR1 mutant mice. *Cell* **79**, 377–388 (1994).
- Chiamulera, C. *et al.* Reinforcing and locomotor stimulant effects of cocaine are absent in mGluR5 null mutant mice. *Nature Neurosci.* **4**, 873–874 (2001).
- Hirono, M., Konishi, S. & Yoshioka, T. Phospholipase C-independent group I metabotropic glutamate receptor-mediated inward current in mouse Purkinje cells. *Biochem. Biophys. Res. Commun.* **251**, 753–758 (1998).
- Knopfel, T., Anchisi, D., Alojado, M. E., Tempia, F. & Strata, P. Elevation of intradendritic sodium concentration mediated by synaptic activation of metabotropic glutamate receptors in cerebellar Purkinje cells. *Eur. J. Neurosci.* **12**, 2199–2204 (2000).
- Canepari, M., Papageorgiou, G., Corrie, J. E. T., Watkins, C. & Ogden, D. The conductance underlying the parallel fibre slow EPSP in rat cerebellar Purkinje neurons studied with photolytic release of L-glutamate. *J. Physiol. (Lond.)* **533**, 765–772 (2001).
- Takechi, H., Eilers, J. & Konnerth, A. A new class of synaptic response involving calcium release in dendritic spines. *Nature* **396**, 757–760 (1998).
- Finch, E. A. & Augustine, G. J. Local calcium signalling by inositol-1,4,5-trisphosphate in Purkinje cell dendrites. *Nature* **396**, 753–756 (1998).
- Miyata, M. *et al.* Deficient long-term synaptic depression in the rostral cerebellum correlated with impaired motor learning in phospholipase C beta4 mutant mice. *Eur. J. Neurosci.* **13**, 1945–1954 (2001).
- Tabata, T., Aiba, A. & Kano, M. Extracellular calcium controls the dynamic range of neuronal metabotropic glutamate receptor responses. *Mol. Cell. Neurosci.* **20**, 56–68 (2002).
- Hartmann, J. *et al.* Gq-dependence of mGluR-mediated synaptic signaling in cerebellar Purkinje cells. *Soc. Neurosci. Abstr.* **339.15** (2002).
- Xiao, B., Tu, J. C. & Worley, P. F. Homer: a link between neural activity and glutamate receptor function. *Curr. Opin. Neurobiol.* **10**, 370–374 (2000).
- Yuan, J. P. *et al.* Homer binds TRPC family channels and is required for gating of TRPC1 by IP3 receptors. *Cell* **114**, 777–789 (2003).
- Montell, C., Birnbaumer, L. & Flockez, V. The TRP channels, a remarkably functional family. *Cell* **108**, 595–598 (2002).
- Strubing, C., Kraplinsky, G., Kraplinsky, L. & Clapham, D. E. TRPC1 and TRPC5 form a novel cation channel in mammalian brain. *Neuron* **29**, 645–655 (2001).
- Gee, C. E., Benquet, P. & Gerber, U. Group I metabotropic glutamate receptors activate a calcium-sensitive transient receptor potential-like conductance in rat hippocampus. *J. Physiol. (Lond.)* **546**, 655–664 (2003).
- Tozzi, A. *et al.* Involvement of transient receptor potential-like channels in responses to mGluR1 activation in midbrain dopamine neurons. *Eur. J. Neurosci.* **18**, 2133–2145 (2003).
- Hofmann, T., Schaefer, M., Schultz, G. & Gudermann, T. Subunit composition of mammalian transient receptor potential channels in living cells. *Proc. Natl Acad. Sci. USA* **99**, 7461–7466 (2002).
- Maejima, T., Hashimoto, K., Yoshida, T., Aiba, A. & Kano, M. Presynaptic inhibition caused by a retrograde signal from metabotropic to cannabinoid receptors. *Neuron* **31**, 463–475 (2001).
- Petralia, R. S. *et al.* Glutamate receptor targeting in the postsynaptic spine involves mechanisms that are independent of myosin Va. *Eur. J. Neurosci.* **13**, 1722–1732 (2001).
- Narasimhan, K., Pessah, I. N. & Linden, D. J. Inositol-1,4,5-trisphosphate receptor-mediated Ca mobilization is not required for cerebellar long-term depression in reduced preparations. *J. Neurophysiol.* **80**, 2963–2974 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Bock for technical assistance, C. Montell and T. Hofmann for advice on TRP channels, and Y.-X. Wang for help with electron microscopy immunohistochemistry. This work was supported by USPHS grants to P.F.W. and D.J.L., and by the Develbiss Fund (D.J.L.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.J.L. (dlinden@jhmi.edu) or P.F.W. (pforley@jhmi.edu).

Testis determination requires insulin receptor family function in mice

Serge Nef^{1*†}, Sunita Verma-Kurvari^{1*}, Jussi Merenmies^{1†}, Jean-Dominique Vassalli², Argiris Efstratiadis³, Domenico Accili⁴ & Luis F. Parada¹

¹Center for Developmental Biology, University of Texas, Southwestern Medical Center, 6000 Harry Hines Boulevard, Dallas, Texas 75390-9133, USA

²University of Geneva, Department of Morphology, Centre Médical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland

³Department of Genetics and Development, and ⁴Naomi Berrie Diabetes Center & Department of Medicine, College of Physicians & Surgeons of Columbia University, New York, New York 10032, USA

* These authors contributed equally to this work

† Present Addresses: University of Geneva, Department of Morphology, Centre Médical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland (S.N.); University of Helsinki, Hospital for Children and Adolescents, Pediatric Nephrology Unit, Developmental and Reproduction Biology, Biomedicum Helsinki, Room B526b, PO Box 700 (Haartmanninkatu 8), FIN-00029 HUS, Finland (J.M.)

In mice, gonads are formed shortly before embryonic day 10.5 by the thickening of the mesonephros and consist of somatic cells and migratory primordial germ cells¹. The male sex-determining process is set in motion by the sex-determining region of the Y chromosome (*Sry*), which triggers differentiation of the Sertoli cell lineage. In turn, Sertoli cells function as organizing centres and direct differentiation of the testis. In the absence of *Sry* expression, neither XX nor XY gonads develop testes², and alterations in *Sry* expression are often associated with abnormal sexual differentiation^{3–8}. The molecular signalling mechanisms by which *Sry* specifies the male pathway and models the undifferentiated gonad are unknown. Here we show that the insulin receptor tyrosine kinase family, comprising *Ir*, *Igf1r* and *Irr*, is required for the appearance of male gonads and thus for male sexual differentiation. XY mice that are mutant for all three receptors develop ovaries and show a completely female phenotype. Reduced expression of both *Sry* and the early testis-specific marker *Sox9* indicates that the insulin signalling pathway is required for male sex determination.

The insulin family signalling pathway is essential for normal growth and development in mice and comprises *Ir*, *Igf1r* and *Irr*. In mice, both *Ir* and *Igf1r* are necessary for pre- and postnatal growth and for development. *Ir* mutant pups die from ketoacidosis within 4 d of birth^{9,10}. *Igf1r* mutant mice die at birth because of respiratory failure and show several abnormalities including reduced size¹¹. Mice lacking both *Ir* and *Igf1r* are even more severely affected¹². By contrast, mice mutant for *Irr* are viable and do not show overt deficiencies, and so far no physiological function has been attributed to this receptor¹³. The identification of *Ir*–*Igf1r* and *Ir*–*Irr* hybrid receptors suggests that there is additional complexity in the signalling potential of this family (see refs 14, 15, and references therein). By using mice with mutations at each of the known genes of insulin receptor tyrosine kinase family members (*Ir*, *Igf1r* and *Irr*), here we provide *in vivo* evidence that insulin receptor family

function is a precondition for the induction of testicular differentiation by *Sry*-dependent processes.

By scanning electron microscopy¹⁶, we analysed the developing genito-urinary system at embryonic day 18.5 (E18.5) in male and female mice that were singly mutant for *Ir*, *Irr* and *Igf1r*, as well as doubly or triply mutant for each of the possible combinations. Normal development of the genito-urinary system, including normal intra-abdominal testicular descent, was observed in wild-type and mutant male mice lacking *Ir*, *Irr*, *Igf1r*, *Ir* and *Irr*, or *Igf1r* and *Irr* (Supplementary Fig. 1 and data not shown). By contrast, no *Ir* *Irr* *Igf1r* triple mutant embryos ($n = 15$) showed a male phenotype, as determined by the presence of descended testes (Supplementary Fig. 1), whereas some *Ir* *Igf1r* double and compound mutant embryos had partially descended gonads (see below). To rule out selective death of triple mutant XY embryos, the presence of Y chromosome was validated by karyotypic ($n = 1$) and molecular ($n = 4$) analysis for the presence of Y-chromosome-specific markers (*Sry*, *Zfy* and *Ubey*; Supplementary Fig. 2).

We carried out histological analysis at E17.5, when the male and female genito-urinary systems are well developed. Similar to wild-type embryos, single and various double mutant embryos showed sexually dimorphic histology (except for the *Ir* *Igf1r* double mutant, see below). In all cases, XX embryos had gonads adjacent to the lower pole of the kidney (Supplementary Fig. 3), whereas the XY gonads were significantly larger and lay at a more caudal position adjacent to the bladder. Figure 1 shows, with exception of the triple mutant, the larger size of male (XY) gonads and the presence of seminiferous tubules. By contrast, the XY triple mutant gonad, in addition to being situated adjacent to the kidney, was histologically indistinguishable from XX gonads with no evidence of seminiferous tubules (Fig. 1h, left and right).

To determine whether morphological sex reversal in XY triple mutants was reflected accurately at the molecular level, markers specific for male (*P450sc*, *Ins13*, *Dhh*, *Mis* and *Sox9*) and female (*Wnt4* and *Figα*) gonads were examined by *in situ* hybridization^{17–19}. None of the testis-specific genes was expressed in triple mutant XY gonads (Fig. 2). By contrast, the expression of female-specific genes in triple mutant XY gonads was indistinguishable from that in normal or mutant XX gonads. Thus, in addition to phenotypic conversion, the XY gonads of triple mutant embryos express the molecular profiles of normally developing female gonads.

In development, the testicular hormones *Mis* and testosterone induce regression of the presumptive female tracts (Müllerian ducts) and stabilization of the Wolffian ducts, respectively. To evaluate the consequences of testicular differentiation impairment in XY triple mutant embryos, we analysed the development of the two reproductive tracts. The genital ducts of *Ir* *Irr* and *Igf1r* *Irr* double mutants and *Irr* single mutants were sexually dimorphic and indistinguishable from those of wild-type embryos (Supplementary Fig. 4). By contrast, the ducts of all *Ir* *Igf1r* *Irr* triple mutant XY embryos resembled the XX structure and showed molecular profiles characteristic of females, including expression of *Wnt4*, *Mis receptor type II* (*MisRII*) and *Wnt7a*²⁰ (Supplementary Fig. 4). These data are consistent with the absence of Sertoli and Leydig cell differentiation.

To examine when insulin receptor family signalling loss affected testicular differentiation, we assessed the presence of early markers including *Sox9* and *Mis*. Expression of *Sox9* is male-specific at E12.5 and is both necessary and sufficient for male sexual determination²¹. At E12.5, XY *Ir* *Igf1r* *Irr* triple mutant mice (XY, $n = 5$; XX, $n = 3$) showed substantially less *Sox9* expression than did XY control mice (Fig. 3a, b). Although small amounts of *Sox9* transcripts were detectable in triple mutant XY genital ridges at E12.5, these were apparently insufficient to activate expression of *Mis* (Fig. 3e). In addition, E12.5 triple mutant XY gonads did not express *P450sc*, but upregulation of female-specific *Wnt4* expression was observed ($n = 3$; data not shown). At E13.5 (XY, $n = 3$) no male-specific markers were detected, but at E14.5 (XY, $n = 1$) expression of

female marker *Figα* was observed (*Sox9*, *Mis* and *P450scs*; Fig. 3g, j, h, k, and data not shown). These results indicate that in the absence of insulin family receptors, male sex determination is blocked early on in gonad differentiation.

Given the reduction or absence of early male-specific markers, we next examined *Sry* expression in the absence of *Ir*, *Igf1r* and *Irr* function. *Sry* gene expression was assessed by real-time polymerase chain reaction with reverse transcription (RT-PCR) analysis in embryos at the tail somite (t.s.) stages (Fig. 3m). Analysis of control mice showed a classical *Sry* expression curve with a peak at the 18–21-t.s. stage (~E11.5). By contrast, expression of *Sry* was reduced in triple mutant mice. These results indicate that insulin signalling is required for normal *Sry* expression and, thus, for initiating the male determination pathway.

A reduction of *Sry* expression in *Ir Igf1r Irr* triple mutant XY gonads might be caused either by a defect in proliferation, resulting in the formation of fewer *Sry*-expressing Sertoli cell precursors, or by reduced *Sry* expression on a per-cell basis. To differentiate between these two possibilities, we used 5-bromodeoxyuridine (BrdU) labelling at E12.5 to examine proliferation in XY gonads lacking the insulin signalling pathway. Subepithelial proliferation in *Ir Igf1r Irr* mutant XY gonads was 80% of that in control testes but similar to that in control ovaries (Supplementary Fig. 5a). A higher proportion of BrdU-labelled germ cells was present close to the coelomic epithelium in both triple mutant XY gonads and control ovaries than in control testes. Excluding germ cells, proliferation of somatic cells in mutant XY gonads was only 72% of that of control testes, which is consistent with the idea that the insulin signalling

pathway may mediate proliferation in the early differentiating testis. These results are in keeping with the reduced size of the triple mutant XY gonad at E12.5 (Fig. 3) and underscore the importance of proliferation in the formation of the male gonad²². Thus, reduced *Sry* levels and a reduction of *Sox9* expression in XY genital ridges suggest that insulin family signalling, either directly or indirectly, functions upstream of *Sry*.

We also found defective testicular differentiation in several compound mutants at the *Ir*, *Igf1r* and *Irr* loci. Both *Ir Igf1r* double mutants and the compound heterozygous *Ir^{+/-} Igf1r^{-/-} Irr^{-/-}* mutant showed partial male-to-female sex reversal (Fig. 4). Analysis of these mutants showed that the XY embryos had varying degrees of gonad descent, with size resembling that of an ovary rather than the generally larger testis (data not shown). Histological examination of the *Ir Igf1r* double mutant XY gonads showed a partial testicular morphology with evidence of seminiferous tubules (Fig. 4a, b). *In situ* hybridization for male- and female-specific markers showed the presence of ovotestes in 66% of the double mutant XY gonads (Fig. 4). In the compound heterozygous *Ir^{+/-} Igf1r^{-/-} Irr^{-/-}* mutant XY gonads, we observed a range of differentiation, including normal gonads, ovotestes and complete sex reversal.

These results suggest that there is a functional redundancy among the three insulin family receptors, *Igf1r*, *Ir* and *Irr*, but with varying relative degrees of contribution. In support of this, we observed sex-reversed phenotypes only when both *Igf1r* alleles were mutated. Male embryos with one mutant allele for *Ir* (*Ir^{+/-} Igf1r^{-/-} Irr^{-/-}*) or no mutant alleles for *Irr* (*Ir^{-/-} Igf1r^{-/-} Irr^{+/-}*) showed a partially sex-reversed phenotype. These data further support the

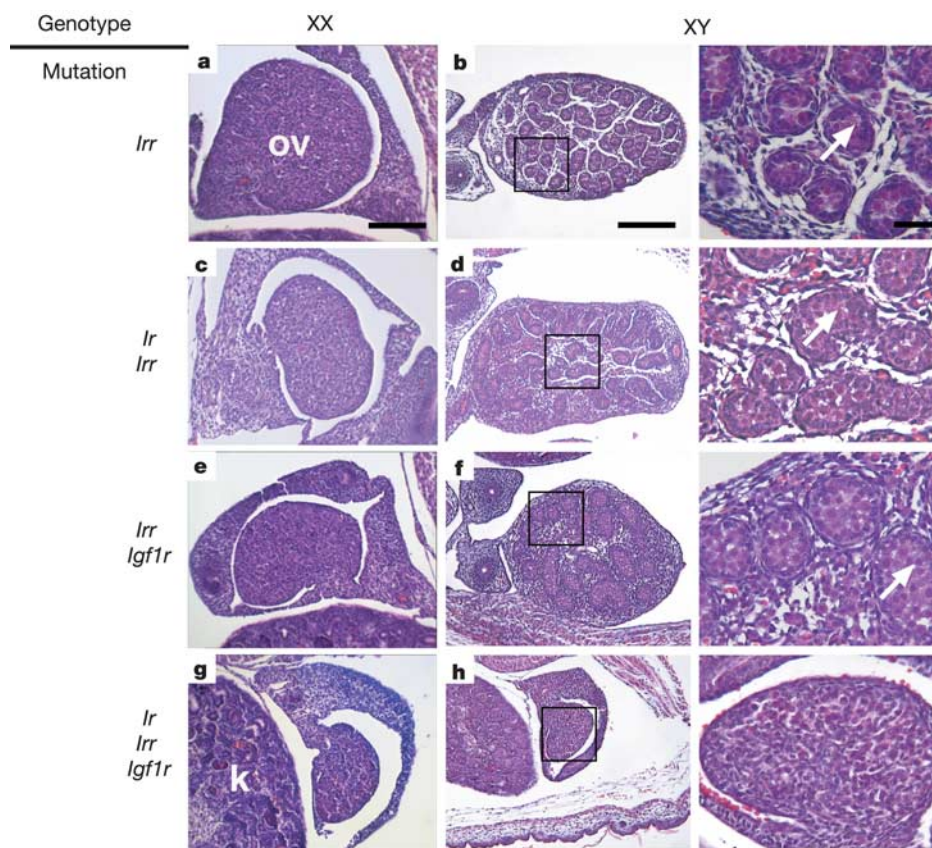


Figure 1 Mice lacking *Ir*, *Irr* and *Igf1r* show complete gonadal sex reversal at the histological level. Shown is haematoxylin and eosin staining of XX gonads (**a**, **c**, **e**, **g**) and XY gonads (**b**, **d**, **f**, **h**; right panels show enlargements of the boxed regions indicated on the left) from single, double and triple mutant embryos. Distinct seminiferous tubules are present in the testes of single and double mutant embryos. Similar to XX gonads of wild

type, single and double mutant combinations, the *Ir Igf1r Irr* triple mutant XY gonads are encapsulated and appear morphologically female with no signs of differentiation (compare **h** with **a**, **c**, **e**, **g**). k, kidney; ov, ovary; t, testis. Scale bars, 200 μ m (**a**, **c**, **e**, **g**), 100 μ m (**b**, **d**, **f**, **h**, left) and 40 μ m (**b**, **d**, **f**, **h**, right).

idea that a threshold of insulin family signalling is required to mediate normal male gonad differentiation. The overall contribution of *Igf1r* is crucial and higher than that of *Ir*, which is itself higher than *Irr*. RT-PCR and *in situ* hybridization showed that all three receptors are expressed in similar quantities in both male and female gonads, although we consistently detected smaller relative amounts of *Irr* in genital ridges (Supplementary Fig. 5b, c). Therefore, sexual dimorphism mediated by the insulin signalling pathway must depend on differential expression or activity of ligands or downstream signalling effectors. The requisite loss of *Irr* to bring about complete sex reversal identifies, to our knowledge, the first physiological function attributable to this receptor.

Studies of XX–XY chimaeric mice have suggested that as little as 20% of Sry-expressing Sertoli precursor cells can direct testicular development. Less than 20% of XY cells permit differentiation into ovaries (reviewed in ref. 23). Our study is at variance with these observations. In compound heterozygous and *Ir Igf1r* mutant mice, often more than 50% of the gonad was composed of testicular tissue and the rest resembled ovarian tissue and contained female germ cells. These data suggest that the presence of 20% XY cells and/or Sertoli cells is not necessarily sufficient for the formation of pure testicular tissue and that other pathways, such as those mediated by insulin or fibroblast growth factor²⁴ signalling, must also have a role.

Sex determination switches are diverse in the animal kingdom and can range from the presence of a Y or an X chromosome, to autosome ratio, to environmental factors, and to social factors (reviewed in refs 1, 25). Despite this great variety in sex determination switches, the basic structure of testes is highly conserved among vertebrates, with similarities found even in *Drosophila*¹. This

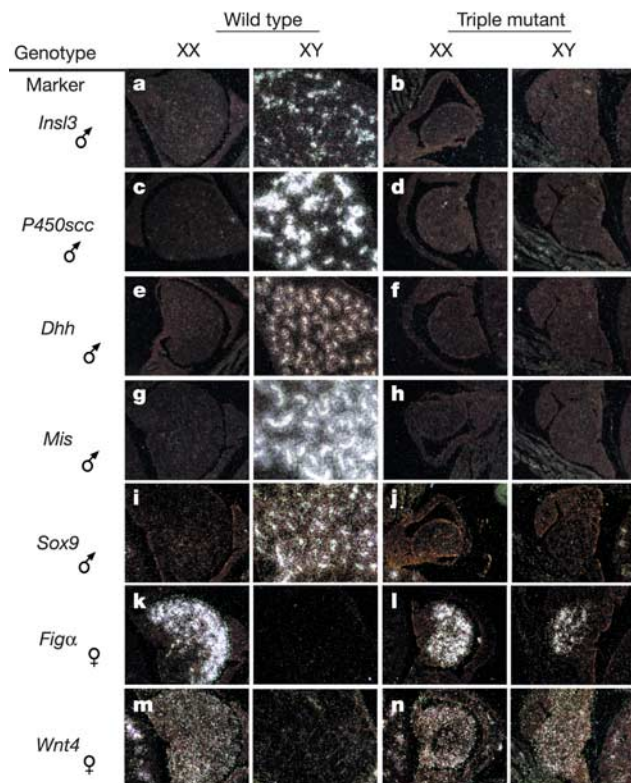


Figure 2 Mice lacking *Ir*, *Irr* and *Igf1r* show complete gonadal sex reversal at the molecular level. Male-specific markers for Leydig cells (*Ins13*, *P450scc*; **a–d**), and Sertoli cells (*Dhh*, *Mis*, *Sox9*; **e–j**) and female-specific markers (*Figα*, *Wnt4*; **k–n**) were used for the molecular characterization of XX gonads (left panels) and XY gonads (right panels). No male-specific markers are present in XY gonads from the triple mutant embryos. Instead, the triple mutant XY gonads express ovarian markers.

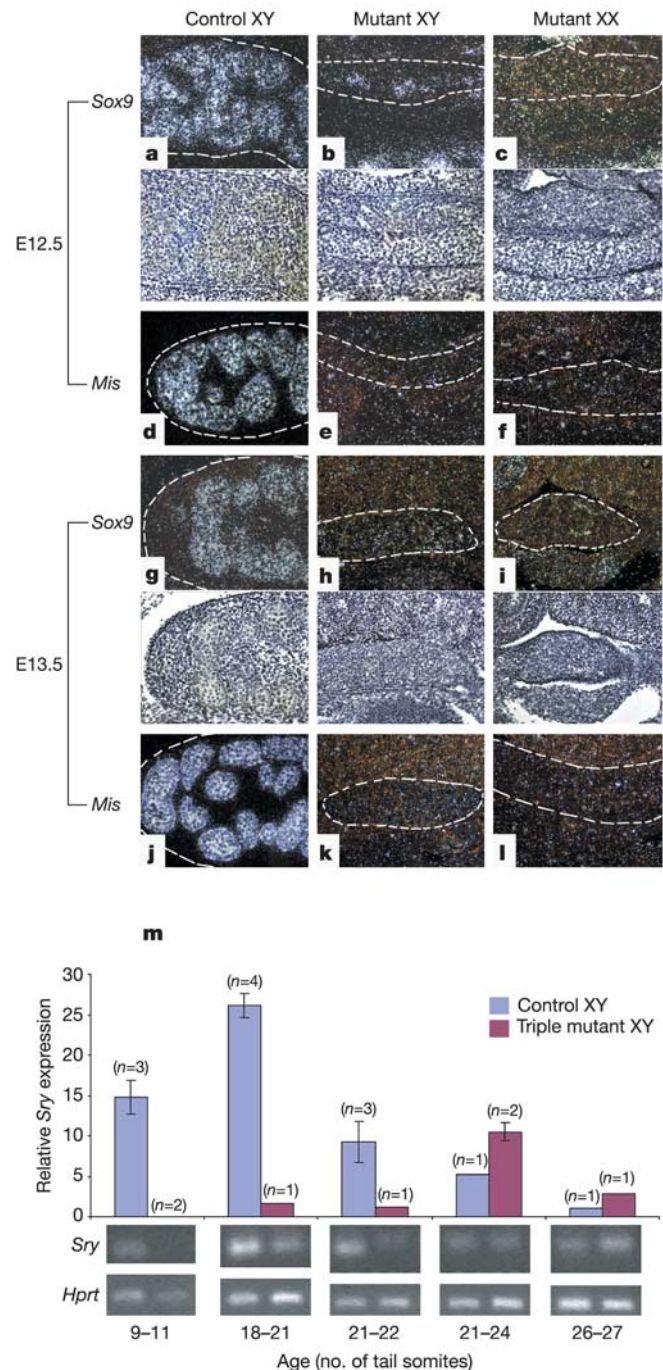


Figure 3 Insulin family signalling is required for early male gonad differentiation. **a–l**, Testis development is blocked in early stages of gonad development. Shown are *in situ* hybridization for *Sox9* (**a–c**, **g–i**, top panels), and the corresponding haematoxylin stain (**a–c**, **g–i**, bottom panels), and *in situ* hybridization for *Mis* (**d–f**, **j–l**) in E12.5 and E13.5 gonads. Robust expression of *Sox9* is observed in control testis at E12.5 (**a**) and E13.5 (**g**), whereas no expression is observed in triple mutant XX gonads (**c**, **i**). By contrast, expression of *Sox9* is reduced in E12.5 (**b**) or absent at E13.5 (**h**) in triple mutant XY gonads. There is no *Mis* expression at either age (**e**, **k**). Note that *Ir Igf1r Irr* triple mutant mice, including the gonads¹², are reduced in size even at E12.5 (compare XY gonads in **a** and **b**). **m**, Expression of *Sry* is reduced in the triple mutant XY gonads. Relative amounts of *Sry* messenger RNA in urogenital ridges were estimated by SYBR green real-time RT–PCR and normalized to those of *Hprt*. Expression of *Sry* in the control XY gonad at t.s. 26–27 was arbitrarily set as 1. The number of embryos analysed at each stage is indicated. Error bars indicate the s.d. A representative gel from the samples used for real-time PCR analysis is shown under the graph.

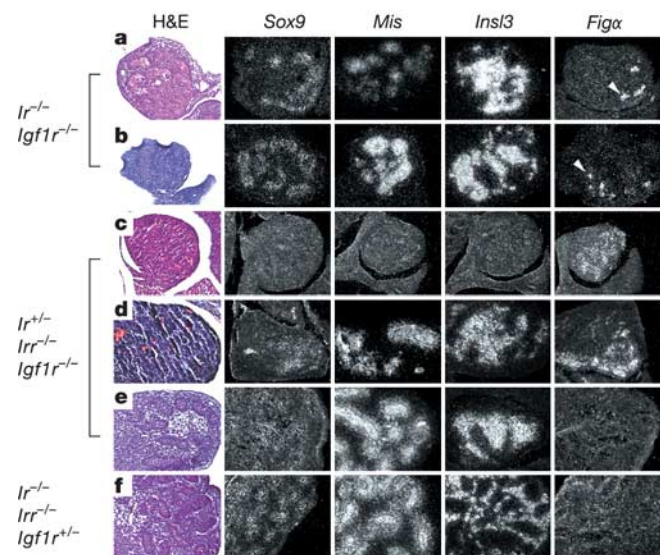


Figure 4 Male-to-female sex reversal phenotype is variable in *Ir Igf1r* double mutant and in compound heterozygous *Ir^{+/−} Igf1r^{−/−} Irr^{−/−}* mutant mice. Shown is the histological and molecular analysis of XY gonads from two *Ir Igf1r* double (a, b), three compound heterozygous *Ir^{+/−} Igf1r^{−/−} Irr^{−/−}* (c–e) and one compound heterozygous *Ir^{−/−} Igf1r^{+/−} Irr^{−/−}* (f) mutant embryos. E17.5 adjacent sections were used for haematoxylin and eosin (H&E) staining and *in situ* hybridization for male-specific (*Sox9*, *Mis* and *Ins13*) and female-specific (*Figα*) markers. In contrast to the compound heterozygous *Ir^{−/−} Igf1r^{+/−} Irr^{−/−}* mutants, which show a normal testicular phenotype, *Ir Igf1r* double and compound heterozygous *Ir^{+/−} Igf1r^{−/−} Irr^{−/−}* XY mutant gonads show variable sex reversed phenotypes ranging from ovotestes to complete sex reversal (c).

morphological conservation is consistent with the hypothesis that the molecular or cellular pathways that generate testis-specific structures may be conserved among species¹. Accordingly, the insulin signalling pathway is present in *Caenorhabditis elegans*, *Drosophila*, amphibians and higher vertebrates. In light of our results, this pathway becomes a particularly good candidate for a role in testis differentiation in many species. □

Methods

Mouse strains and genotyping

The *Ir*, *Irr* and *Igf1r* mutant mice have been described^{9,11,13}. To generate all combinations of double and triple mutants, mice that were heterozygous at both the *Ir* and the *Igf1r* loci were mated with either wild-type mice or mice that were homozygous null at the *Irr* locus. Mice used for generating the *Ir Igf1r Irr* triple mutant mice were on a C57BL/6J background, whereas mice used for generating the *Ir Igf1r* double mutants were on a 129 background. We genotyped mice by described protocols^{9,11,13}. The presence of the *Sry* PCR product (450 base pairs) was used for the routine sexing of embryos (primers 15869, 5′-CAG CCC TAC AGC CAC ATG AT-3′, and 15870: 5′-GAG TAC AGG TGT GCA GCT CTA-3′).

Microscopy, histology and *in situ* hybridization

E18.5 embryos were processed for scanning electron microscopy as described¹⁶. For histology, embryos were fixed overnight in 4% paraformaldehyde. We mounted 7-μm serial paraffin sections on slides and stained them with haematoxylin and eosin or processed them for *in situ* hybridization¹⁶. ³⁵S-labelled antisense probes were generated from linearized clones of *Ins13*, cholesterol side-chain cleavage cytochrome *P450* (*P450sc*), *Dhh*, *Sox9*, *Mis*, *Figα*, *wnt4*, *MisR1* and *wnt7a*. After hybridization, sections were emulsion-coated, developed, stained with haematoxylin, mounted and photographed by an Olympus microscope.

SYBR green real-time RT-PCR analysis

For *Sry*, total RNAs from triple mutant and littermate control XY gonads were isolated using Trizol (Invitrogen). After DNase treatment, total RNAs were reverse-transcribed and one-fortieth of the complementary DNA template was used for SYBR green real-time PCR. Amplification of *Sry* and *Hprt* were done with the following primers: for *Sry*, 5′-TAC TGG ACT TCT AAG TGG TTA G-3′ and 5′-TTT TTT TTT TTT TTT TTT GGG ATG G-3′ (ref. 26); for *Hprt*, 5′-CCT GCT GGA TTA CAT TAA AGC ACT-3′ and 5′-GTC AAG GGC ATA TCC AAC AAC AAA-3′. Quantitative RT-PCR for each embryo was done in 3–6 replicates and the *Sry* values were normalized to the corresponding amounts of *Hprt*

RNA. We calculated relative amounts of *Sry* by the comparative C_T ($2^{-\Delta\Delta C_T}$) method described in user bulletin no. 2 of the ABI Prism 7700 Sequence Detection System (Applied Biosystems).

For the insulin family receptors, 250 ng of total RNA from gonads at E10.5, E11.5, E12.5 and E13.5, extracted using RNAqueous (Ambion), were reverse-transcribed by using random hexamers. Specific amplifications of *Ir*, *Irr*, *Igf1r* and *Gapdh* were done with SYBR green and the following primers: for *Ir*, 5′-CAA TGG GAC CAC TGT ATG CAT CT-3′ and 5′-GTC CGG CAC GTA CAC AGA AGA-3′; for *Irr*, 5′-CCA GAA AAA ATT TGA AAA CTT CCT ACA C-3′ and 5′-GAG TTT TTT CCT AGT CTG AGA AGT C-3′; for *Igf1r*, 5′-GGA GTG TCC CTC AGG CTT CA-3′ and 5′-CAT CGC CGC AGA CTT TGG-3′; for *Gapdh*, 5′-TCC ATG ACA ACT TTG GCA TTG-3′ and 5′-CAG TCT TCT GGG TGG CAG TGA-3′. Expression of each receptor was normalized to the corresponding amount of *Gapdh* RNA at that age. Relative amounts of each receptor were calculated by 2^{-C_T} (C_T , the cycle number at which the signal reached the threshold of detection). Each C_T value used for the calculation was the mean of results obtained from three embryos analysed in triplicate. Statistical analysis was done with ANOVA.

For non-quantitative RT-PCR analysis of insulin family receptors, we used primers interrupted by an intron: for *Ir*, 5′-GTG AGG CAG AGA CCC GTG TTG CGG TT-3′ and 5′-TCT CTC TGG ACA GTT ATC AGG GGG A-3′; for *Igf1r*, 5′-ATG AAC CCG AAA CCA GAG TGG CCA TC-3′ and 5′-CAT ATC AGG GCA GTT GTC CGG CTT G-3′; for *Irr*, 5′-GAT TCA GCT ACA AGA GCT GAT GAG-3′ and 5′-TTT CTC TCT GAG GGC AGA GCA CTT C-3′.

Cell proliferation assay

Timed pregnant females were injected intraperitoneally with BrdU at a final concentration of 50 mg per kg (body weight) 2 h before being killed. We carried out BrdU immunohistochemistry²⁷ and counted positive cells²⁴ as described. Two to four parasagittal sections of 7 μm from triple mutant XY gonads ($n = 3$), control XY gonads ($n = 3$) and control XX gonads ($n = 4$) at E12.5 were analysed. BrdU-positive cells were counted in an area extending 35 μm from the coelomic epithelium into the gonad. Germ cells were identified by a large round nucleus and the somatic cell count was obtained by subtracting the germ cell count from the total cell count. Because the triple mutant embryos were smaller than their littermates, the final cell counts were adjusted to reflect differences in total gonad area for each genotype. *P* values were obtained by a two-tailed *t*-test.

Received 9 July; accepted 5 September 2003; doi:10.1038/nature02059.

- Capel, B. The battle of the sexes. *Mech. Dev.* **92**, 89–103 (2000).
- Capel, B. *et al.* Deletion of Y chromosome sequences located outside the testis determining region can cause XY female sex reversal. *Nature Genet.* **5**, 301–307 (1993).
- Lee, C. H. & Taketo, T. Normal onset, but prolonged expression, of *Sry* gene in the B6.YDOM sex-reversed mouse gonad. *Dev. Biol.* **165**, 442–452 (1994).
- Washburn, L. L., Albrecht, K. H. & Eicher, E. M. C57BL/6J-T-associated sex reversal in mice is caused by reduced expression of a *Mus domesticus Sry* allele. *Genetics* **158**, 1675–1681 (2001).
- Tevosian, S. G. *et al.* Gonadal differentiation, sex determination and normal *Sry* expression in mice require direct interaction between transcription partners GATA4 and FOG2. *Development* **129**, 4627–4634 (2002).
- Nagamine, C. M., Morohashi, K., Carlisle, C. & Chang, D. K. Sex reversal caused by *Mus musculus domesticus* Y chromosomes linked to variant expression of the testis-determining gene. *Sry. Dev. Biol.* **216**, 182–194 (1999).
- Hammes, A. *et al.* Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation. *Cell* **106**, 319–329 (2001).
- Albrecht, K. H., Young, M., Washburn, L. L. & Eicher, E. M. *Sry* expression level and protein isoform differences play a role in abnormal testis development in C57BL/6J mice carrying certain *Sry* alleles. *Genetics* **164**, 277–288 (2003).
- Accili, D. *et al.* Early neonatal death in mice homozygous for a null allele of the insulin receptor gene. *Nature Genet.* **12**, 106–109 (1996).
- Joshi, R. L. *et al.* Targeted disruption of the insulin receptor gene in the mouse results in neonatal lethality. *EMBO J.* **15**, 1542–1547 (1996).
- Liu, J. P., Baker, J., Perkins, A. S., Robertson, E. J. & Efstratiadis, A. Mice carrying null mutations of the genes encoding insulin-like growth factor I (*Igf-1*) and type 1 IGF receptor (*Igf1r*). *Cell* **75**, 59–72 (1993).
- Louvi, A., Accili, D. & Efstratiadis, A. Growth-promoting interaction of IGF-II with the insulin receptor during mouse embryonic development. *Dev. Biol.* **189**, 33–48 (1997).
- Kitamura, T. *et al.* Preserved pancreatic β-cell development and function in mice lacking the insulin receptor-related receptor. *Mol. Cell. Biol.* **21**, 5624–5630 (2001).
- Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **103**, 211–225 (2000).
- Fernandez, A. M. *et al.* Functional inactivation of the IGF-I and insulin receptors in skeletal muscle causes type 2 diabetes. *Genes Dev.* **15**, 1926–1934 (2001).
- Nef, S. & Parada, L. F. Cryptorchidism in mice mutant for *Ins13*. *Nature Genet.* **22**, 295–299 (1999).
- Swain, A. & Lovell-Badge, R. Mammalian sex determination: a molecular drama. *Genes Dev.* **13**, 755–767 (1999).
- Vainio, S., Heikkilä, M., Kispert, A., Chin, N. & McMahon, A. P. Female development in mammals is regulated by Wnt-4 signalling. *Nature* **397**, 405–409 (1999).
- Soyal, S. M., Amleh, A. & Dean, J. *Figα*, a germ cell-specific transcription factor required for ovarian follicle formation. *Development* **127**, 4645–4654 (2000).
- Parr, B. A. & McMahon, A. P. Sexually dimorphic development of the mammalian reproductive tract requires Wnt-7a. *Nature* **395**, 707–710 (1998).
- Clarkson, M. J. & Harley, V. R. Sex with two SOX on: *SRY* and *SOX9* in testis development. *Trends Endocrinol. Metab.* **13**, 106–111 (2002).
- Schmah, J. & Capel, B. Cell proliferation is necessary for the determination of male fate in the gonad. *Dev. Biol.* **258**, 264–276 (2003).
- Burgoyne, P. & Palmer, S. in *Gonadal Development and Function* (ed. Hiller, S. G.) 17–29 (Raven, New York, 1993).

24. Colvin, J. S., Green, R. P., Schmah, J., Capel, B. & Ornitz, D. M. Male-to-female sex reversal in mice lacking fibroblast growth factor 9. *Cell* **104**, 875–889 (2001).
25. Scherer, G. Introduction: vertebrate sex determination and gonadal differentiation. *Cell. Mol. Life Sci.* **55**, 991–909 (1999).
26. Jeske, Y. W., Bowles, J., Greenfield, A. & Koopman, P. Expression of a linear Sry transcript in the mouse genital ridge. *Nature Genet.* **10**, 480–482 (1995).
27. Kerner, S. G., Erwin, T. M. & Parada, L. F. Brain remodeling due to neuronal and astrocytic proliferation after controlled cortical injury in mice. *J. Neurosci Res.* **66**, 317–326 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Menon, Y.-J. Li, S. McKinnon, Z. Jorai, J. Kargel, T. Shipman, J.-L. Pitetti and the genomic platform of the Frontiers in Genetics at the National Center of Competence in Research for technical assistance; all members of the Parada laboratory for discussions; A. McMahon, R. Behringer, K. Parker, R. Lovell-Badge and J. Dean for probes; and J. Wilson, K. Parker, J. Graff, J. Goldstein and G. Karsenty for critically reading the manuscript. This work was funded by an Excellence in Education Endowment to L.F.P.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.F.P. (luis.parada@utsouthwestern.edu).

Eyes absent represents a class of protein tyrosine phosphatases

Jayanagendra P. Rayapureddi¹, Chandramohan Kattamuri¹, Brian D. Steinmetz¹, Benjamin J. Frankfort^{2,3}, Edwin J. Ostrin^{2,3}, Graeme Mardon^{2,4} & Rashmi S. Hegde¹

¹Division of Developmental Biology, Cincinnati Childrens Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA

²Departments of Molecular and Human Genetics, ³Pathology, and

⁴Ophthalmology, Neuroscience and Program in Developmental Biology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

The Eyes absent proteins are members of a conserved regulatory network implicated in the development of the eye, muscle, kidney and ear^{1–7}. Mutations in the Eyes absent genes have been associated with several congenital disorders including the multi-organ disease bronchio-oto-renal syndrome⁸, congenital cataracts⁹ and late-onset deafness¹⁰. On the basis of previous analyses it has been shown that Eyes absent is a nuclear transcription factor, acting through interaction with homeodomain-containing Sine oculis (also known as Six) proteins¹¹. Here we show that Eyes absent is also a protein tyrosine phosphatase. It does not resemble the classical tyrosine phosphatases that use cysteine as a nucleophile and proceed by means of a thiol-phosphate intermediate¹². Rather, Eyes absent is the prototype for a class of protein tyrosine phosphatases that use a nucleophilic aspartic acid in a metal-dependent reaction. Furthermore, the phosphatase activity of Eyes absent contributes to its ability to induce eye formation in *Drosophila*.

The Eyes absent proteins contain a divergent 200–300-residue, amino-terminal region and a conserved carboxy-terminal domain of 271–274 residues (Eya domain (ED)), which is critical for activity and believed to participate in protein–protein interactions^{13–15}. Plant Eyes absent proteins consist of the ED with a 16–21-residue N-terminal extension. Although Eyes absent does not have significant sequence homology with any known structural or functional family of proteins, the ED contains signature motifs of the haloacid dehalogenase (HAD) hydrolases, prompting us to examine whether Eyes absent has enzymatic activity. Members of the HAD superfamily, which includes several types of enzyme including phosphatases, are characterized by three short motifs (Fig. 1a)¹⁶. Motif I at

the N-terminal end of the domain consists of DXDXT/S. Motif II is hhh(T/S) where 'h' is a hydrophobic residue. Motif III, K(X_n)-hhhhGDXXX(D/E), is less conserved and is found at the C-terminal end of the domain.

Bacterially generated Eyes absent Eya domains from mouse (Eya1–3) and *Drosophila melanogaster* (Eya), as well as full-length *Arabidopsis thaliana* EYA, were able to hydrolyse the standard phosphatase substrate p-nitrophenylphosphate (pNPP) in the presence of magnesium (Fig. 1b; see also Supplementary Fig. 1). Eya3(ED) was the most potent of the mouse phosphatases followed by Eya2(ED) and Eya1(ED). *Drosophila* Eya(ED) was the least potent. The *A. thaliana* protein displayed nearly 3.5-times greater phosphatase activity than Eya3(ED) and is the strongest Eyes absent phosphatase that we have tested so far. Divalent cations varied in their ability to promote the reaction of Eya3(ED) (Fig. 1c), with Mg²⁺ resulting in the largest K_{cat} (catalytic rate constant), whereas binding affinity was higher for Ni²⁺, Co²⁺ and Mn²⁺. No activity was detected in the presence of Ca²⁺ or Zn²⁺. The variation in the catalytic activity of the various Eyes absent proteins might reflect both intrinsic properties as well as the effect of varying percentages of active enzyme obtained when the proteins are overexpressed in bacteria. The phosphatase activity of the Eyes absent proteins is in the range of the well characterized, potent protein tyrosine phosphatases (PTPs), which we assayed in parallel experiments using pNPP as a substrate; *A. thaliana* Eyes absent (K_{cat} of 3,690 s^{−1}) is 10% as active as PTP1B (K_{cat} of 36,000 s^{−1}) and 1.3 times as active as full-length SHP1 (K_{cat} of 2,785 s^{−1}).

By analogy with the HAD enzymes^{17–19}, dephosphorylation catalysed by Eyes absent probably involves two steps (proposed reaction mechanism shown in Supplementary Fig. 2). First, a nucleophilic attack by the invariant first Asp in motif I results in a covalent phospho-enzyme intermediate, with the second Asp in motif I possibly serving as a general acid by donating a proton to the leaving group. The Thr/Ser residue in motif I is expected to orient the nucleophilic Asp side chain for the reaction. The transition state would be stabilized by an interaction with the positively charged cavity comprising Lys (motif III), Mg²⁺, the Ser/Thr side chain of motif II, and other groups around the phosphate-binding site. An activated water molecule then hydrolyses the phospho-enzyme intermediate, releasing inorganic phosphate and regenerating the active enzyme. In the various HAD family enzymes it has been speculated that the role of activating the attacking water molecule is carried out by the second Asp in motif I²⁰, the lysine residue in motif III²¹, or other as yet unidentified active site residues. To test this proposed mechanism, as well as to verify the alignment in Fig. 1a, we generated and assayed a series of mutationally altered forms of Eya3(ED), targeting putative active site residues (Supplementary Table 1). Conservative replacement of motif I residues—the nucleophilic Asp 246 (to Asn), the general acid/base Asp 248 (to Asn) and the nucleophile-orienting Thr 250 (to Ala)—inactivated the enzyme (<1% activity). Replacement of the nucleophilic aspartate in *Drosophila* Eya(ED) (Asp 493) and *A. thaliana* EYA (Asp 25) also

Table 1 **Substrate specificity of Eyes absent**

Substrate	K_m (mM)	K_{cat} (s ^{−1})
Phosphotyrosine	1.473 (±0.39)	124.8 (±13.6)
Phosphoserine	No detectable activity	–
Phosphothreonine	No detectable activity	–
RRLEDAEpYAARG	0.247 (±0.003)	97.7 (±1)
ENDpYINSL	0.596 (±0.1)	31.52 (±2.38)
DADEpYLIPOQG	0.488 (±0.09)	23.02 (±1.7)
TRDlpYETDYRK	0.123 (±0.02)	55.44 (±6.8)
RRApTVA	No detectable activity	–
RRApSVA	No detectable activity	–
Phospholipids	No detectable activity	–
dATP, dTTP, dCTP, dGTP	No detectable activity	–

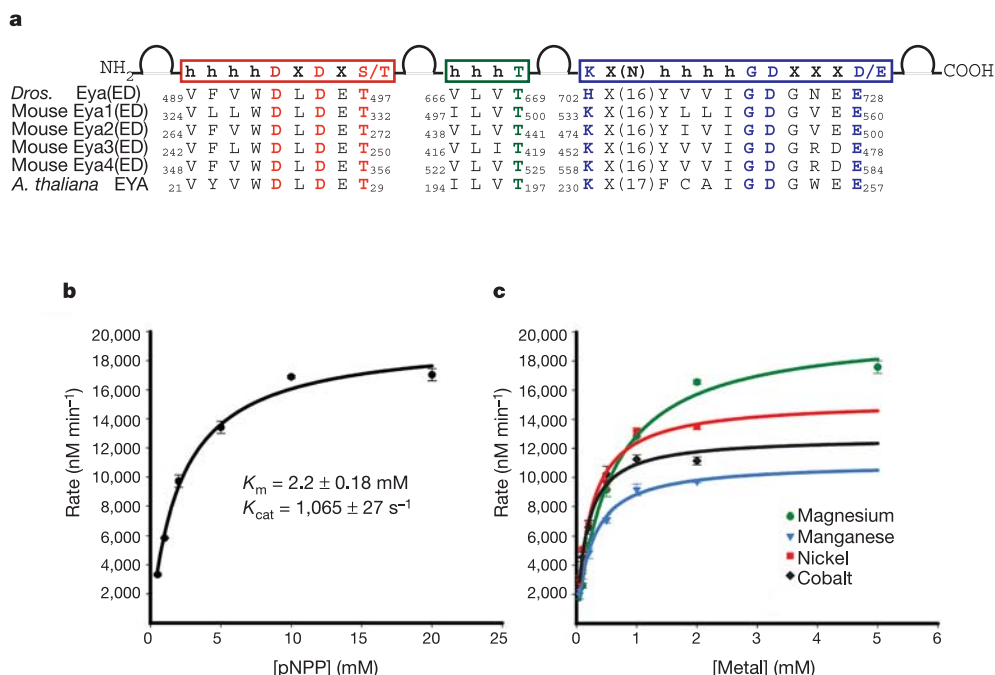


Figure 1 Eyes absent is a phosphatase. **a**, Schematic representation of the motifs common to the HAD family of enzymes. Motifs I, II and III are shown in red, green and blue boxes, respectively. The corresponding sequences in mouse Eya1, -2, -3, -4, *Drosophila* Eya and *A. thaliana* EYA are shown. **b**, Saturation kinetics of Eya3(ED). Rate of hydrolysis of substrate (pNPP) (assay buffer 20 mM MES pH 6, 2 mM MgCl₂) is plotted against increasing substrate concentration. Data were fitted to the Michaelis-Menten equation,

and the resulting kinetic parameters are indicated. **c**, Rate of substrate hydrolysis by Eya3(ED) is plotted as a function of metal concentration at saturating pNPP concentration. The metals behave like pseudo-substrates allowing K_{cat} and K_{metal} to be determined. Mg²⁺ $K_{cat} = 1,099$ s⁻¹, $K_{metal} = 0.57$ mM; Co²⁺ $K_{cat} = 694$ s⁻¹, $K_{metal} = 0.17$ mM; Ni²⁺ $K_{cat} = 830$ s⁻¹, $K_{metal} = 0.23$ mM; Mn²⁺ $K_{cat} = 599$ s⁻¹, $K_{metal} = 0.25$ mM.

led to inactivation. Replacing motif II Thr 419—which is expected to participate in substrate binding and transition state stabilization—with an alanine led to a 24-fold drop in enzymatic activity of Eya3(ED). Replacement of motif III residues Gly 473 (with Ala) and Glu 478 (with Gln) also inactivated the enzyme, consistent with their proposed structural (Gly 473) and metal coordination (Glu 478) roles. Notably, replacing the other metal-coordinating residue in motif III, Asp 474, with an asparagine only resulted in a 70% reduction in activity, perhaps reflecting a relatively neutral, sub-optimum cation-binding site where asparagine can participate in magnesium coordination.

In order to identify the preferred substrates of the Eyes absent phosphatases we assayed various phosphorylated non-peptides and peptides (Table 1). Eya3(ED) efficiently hydrolysed phosphotyrosine, while showing no activity towards phosphothreonine and phosphoserine. It was also inactive against nucleotide triphosphates. Corroborating the observed preference for phosphotyrosine, a clear preference for phosphotyrosine-containing peptides over phosphothreonine- or phosphoserine-containing peptides was apparent. The phosphopeptides tested included DADEpYLIPQQG (auto-phosphorylation site in the epidermal growth factor receptor²²), ENDpYINASL (highly conserved region of the T-cell phosphatase TC-PTP), RRLIEDAepYAARG (Tyr 419 phosphorylation site in Src), TRDIpYETDYYRK (phosphorylation site in the insulin receptor) and RRA(pT/S)VA (standard peptides for distinguishing Ser/Thr protein phosphatases from acidic and alkaline phosphatases²³). Affinity for peptide substrates was consistently greater than for phosphotyrosine, and the specificity constant for the Src- and insulin receptor peptides is significantly greater than for phosphotyrosine, suggesting that peptides may be the preferred targets of Eyes absent activity. In parallel experiments using the Src peptide (Supplementary Table 2), *A. thaliana* EYA (K_{cat} of 259 s⁻¹) retained the greatest activity among the Eyes absent proteins, with Eya2(ED)

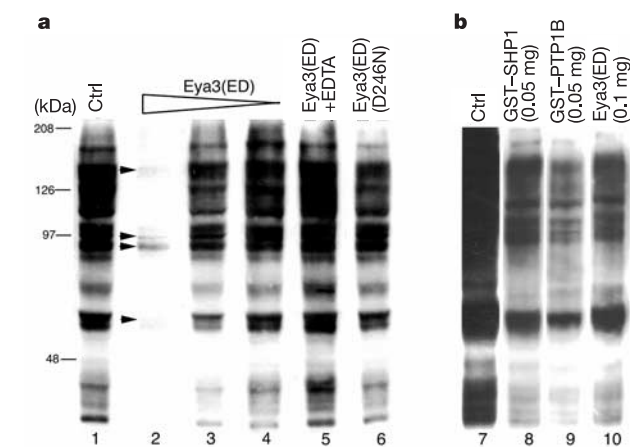


Figure 2 Eya3(ED) can dephosphorylate proteins in a dose- and metal-dependent fashion. **a**, Pervanadate-stimulated 293 cell lysates were electrophoresed on SDS-PAGE gels, transferred onto PVDF membranes and incubated with Eya3(ED) (lanes 2–4; 1,500 μ g, 150 μ g and 15 μ g), 1,500 μ g Eya3(ED) plus 20 mM EDTA (lane 5) or 1,500 μ g Eya3(ED)(D246N) (lane 6). Residual phosphotyrosine-containing bands were visualized by western blots using anti-phosphotyrosine antibody. Arrows indicate phosphorylated proteins that appear to be resistant to Eya3(ED) phosphatase activity. **b**, Comparison of the abilities of Eya3(ED) (lane 10), SHP1 (lane 8) and PTP1B (lane 9) to dephosphorylate proteins in a blot assay, performed as described in **a**. The lanes show membranes incubated with control buffer (lane 7), 50 μ g GST-SHP1 (lane 8), 50 μ g GST-PTP1B (lane 9) and 100 μ g Eya3(ED) (lane 10).

(K_{cat} of 78.6 s^{-1}) and Eya3(ED) (K_{cat} of 97.7 s^{-1}) having comparable activities. Although the phosphatase activity of the Eyes absent proteins was weaker than that of the potent control phosphatases tested (TC-PTP K_{cat} of $5,850 \text{ s}^{-1}$; GST-PTP1B K_{cat} of $5,626 \text{ s}^{-1}$, full-length SHP1 K_{cat} of 345 s^{-1}), the steady-state parameters for Eyes absent phosphatase activity are comparable to those of other tyrosine phosphatases towards non-cognate substrate peptides^{22,24}.

Eya3(ED) showed no activity towards phospholipid vesicles generated by combining dioleoyl phosphatidylcholine, dioleoyl phosphatidylserine and phosphatidylinositol-3,4,5-triphosphate (in a parallel experiment the lipid phosphatase PTEN had a specific activity of $0.85 \text{ nM min}^{-1} \text{ mg}^{-1}$), eliminating the possibility that it is a phospholipid phosphatase.

Eyes absent proteins were able to dephosphorylate the commonly used model protein substrate myelin basic protein (MBP) (Supplementary Table 2). The most robust activity was observed with *A. thaliana* EYA (dephosphorylation rate of $157 \text{ nM min}^{-1} \text{ mg}^{-1}$ enzyme). Eya3(ED) released $146 \text{ nM PO}_4^{2-} \text{ min}^{-1} \text{ mg}^{-1}$ enzyme. Other Eyes absent proteins tested also dephosphorylated $\gamma^{32}\text{P}$ -labelled MBP with relative efficiencies mirroring the general trends reported above for peptide substrates. In parallel experiments the strong phosphatases PTP1B and TC-PTP were 3.6 and 12.7 times more active than Eya3(ED) respectively, but full-length SHP1 had only 60% of the activity of Eya3(ED). Notably, activity differences among the Eyes absent proteins and control phosphatases towards MBP were less pronounced than with peptide substrates and pNPP.

Eya3(ED) could also dephosphorylate proteins in 293 cell extracts in a dose-dependent manner (Fig. 2a). The activity was metal-dependent, being completely abrogated by EDTA. Furthermore, Eya3(ED) lacking the nucleophilic aspartate (D246N) had significantly attenuated activity. Eya3(ED) exhibits preferences among the protein substrates, suggesting that the catalytic activity is sensitive to the context of the phosphotyrosine group and mirroring the observations with peptide substrates described in Table 1. Thus, in two different assays Eyes absent acted as a protein tyrosine phosphatase. When the protein dephosphorylation abilities of Eya3(ED), SHP1 and PTP1B in 293 cell extracts were compared (Fig. 2b), the level of dephosphorylation with 0.1 mg Eya3(ED) and 0.05 mg SHP1 were approximately equivalent. PTP1B was more effective than either SHP1 or Eya3(ED). In parallel with the trend observed with MBP as a substrate, the difference in activity between Eyes absent, SHP1 and PTP1B is less pronounced with protein substrates than might be expected based on the catalytic parameters using pNPP and peptides as substrates.

Further biochemical evidence for the mechanism and specificity

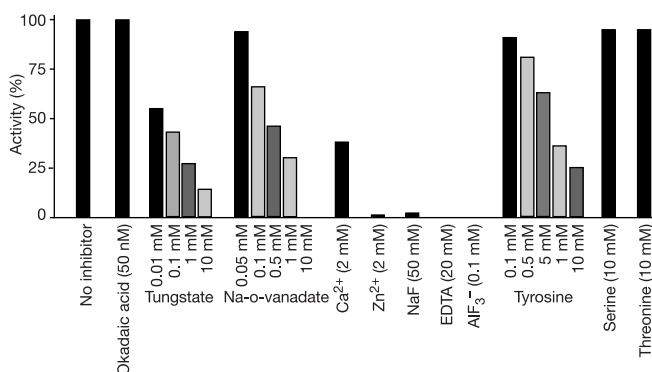


Figure 3 The effect of inhibitors on the phosphatase activity of Eya3(ED). Concentrations of the inhibitors used are indicated on the x axis and the activity relative to that in the absence of any inhibitor is indicated by the bars.

of Eyes absent was derived from its susceptibility to inhibitors (Fig. 3). EDTA and EGTA completely inactivated the enzyme. Ca^{2+} and Zn^{2+} inhibited catalysis, probably by competing with Mg^{2+} . Orthovanadate inactivated Eya3(ED) in a concentration-dependent manner, suggesting that it may mimic phosphate and act as a product inhibitor. Tungstate, another phosphate analogue, also inhibited Eya3(ED) in a dose-dependent fashion. AlF_3^- , which can form transition-state analogues with enzymes containing active site aspartates, inhibited Eya3(ED). Okadaic acid, a Ser/Thr phosphatase inhibitor, did not affect the activity of Eya3(ED). NaF, which is known to inhibit enzymes that require magnesium, inactivated Eya3(ED). Dose-dependent product inhibition by tyrosine was observed, whereas serine and threonine were unable to inhibit the enzyme. These results support the classification of Eyes absent as a metal-dependent tyrosine phosphatase.

To determine whether Eya phosphatase function is required *in vivo*, we expressed *eya*^{D493N} in a *Drosophila* rescue assay using the GAL4/UAS system. *UAS-eya*^{D493N} was expressed in the developing eye under the control of *eyeless*-GAL4 (*ey*-GAL4) in an *eya*² mutant background. As *eya*² flies have no ommatidia but are otherwise viable and fertile²⁵, we compared phosphatase-deficient *eya* and wild-type *eya* with regard to their ability to restore ommatidial development (Fig. 4). We found that the degree of ommatidial restoration was significantly less with *eya*^{D493N} than with wild-type *eya*, confirming our *in vitro* findings that Asp 493 is important for Eya function. These results suggest that the phos-

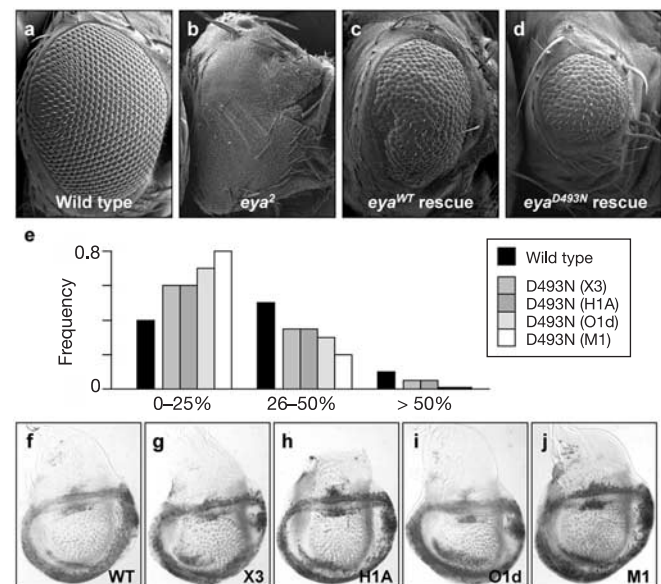


Figure 4 Loss of phosphatase activity compromises Eya function *in vivo*. **a-d**, Scanning electron micrographs of representative adult eyes. **a**, Wild type; **b**, *eya*² homozygote; **c**, *eya*² rescued with wild-type *eya* (*UAS-eya*^{WT}); **d**, *eya*² rescued with phosphatase-deficient *eya* (*UAS-eya*^{D493N}). **e**, Comparison of the degree of *eya*² rescue with *UAS-eya*^{WT} (black, $n = 90$) and four independent *UAS-eya*^{D493N} transgenic lines (X3, $n = 168$; H1A, $n = 210$; O1d, $n = 34$; M1, $n = 56$). Rescued eyes were assigned to one of three classes: 0–25%, 26–50%, or >50% of normal (defined as (observed eye size/wild-type eye size) $\times$ 100). Eye development is rescued to a greater extent with *UAS-eya*^{WT} than with any of the *UAS-eya*^{D493N} transgenic lines. All differences were determined to be significant ($P < 0.05$) with a test of binomial proportions. **f-j**, *UAS-eya*^{WT} and *UAS-eya*^{D493N} were misexpressed in the wing pouch with 30A-GAL4, and late third-instar imaginal discs were dissected and stained with antibody to wild-type Eya protein. Eya^{D493N} (**g-j**) is detected at levels that are equal to or greater than Eya^{WT} (**f**) for all four transgenic lines, suggesting that the rescue results in **a-e** are due to differences in protein function and not levels of transgene expression.

phatase activity of Eyes absent has a role in *Drosophila* eye development. The observed degree of rescue leaves open the possibility that Eya may have additional activities that are important for eye development.

Further supporting the biological importance of the catalytic activity of Eyes absent is a previous genetic analysis²⁶ showing that two mutations in the conserved motifs, *eya*^{T497M} (motif I) and *eya*^{G473E} (motif III), resulted in varying degrees of viability, as well as deficiencies in rescue and ectopic eye development assays. Of particular relevance was the observation that although the active site mutant Eya^{T497M} was able to synergize with Dachshund to induce ectopic eyes, it produced no ectopic eye structures when co-expressed with Sine oculis (So). As wild-type Eya and So act synergistically in ectopic eye formation assays, these results were attributed to a physiologically relevant interaction between Eya and So mediated by the Thr 497-containing region of Eyes absent. In light of the data presented here, these observations are also consistent with the possibility that Sine oculis may be a substrate of Eyes absent with Thr 497 being a catalytically important residue.

Our evidence suggests that the Eyes absent proteins represent a family of protein tyrosine phosphatases, and that Eyes absent is not a simple transcription factor. The biological implication of our data is that Eyes absent may participate directly in the modulation of cellular signalling through its phosphatase activity. Furthermore, based on the phenotype of Eya1 null mice⁷, it is likely that the phosphatase activity is critical for the development of multiple organs. □

Methods

Recombinant proteins

Eya3(ED) (residues 238–510), Eya1(ED) (residues 319–591), Eya2(ED) (residues 262–532) and *Drosophila* Eya(ED) (residues 481–760) were produced as glutathione S-transferase (GST) fusions in pGEX 4T-1. Full-length *A. thaliana* EYA was produced as a His-tagged protein in pET15b. Proteins were purified using glutathione- (Eya3(ED), Eya1(ED), Eya2(ED), Eya(ED)) or Ni-NTA (*A. thaliana* EYA) agarose affinity chromatography. Fusion proteins were treated with thrombin to yield untagged proteins. Eya3, Eya1, Eya and *A. thaliana* EYA proteins were purified to homogeneity by size-exclusion chromatography, whereas Eya2 was not further purified.

Control phosphatases used were GST fusion forms of full-length human PTP1B and full-length human SHPTP1, and a truncated form of human T-cell protein tyrosine phosphatase (11 kDa C-terminal deletion (TC-PTPΔC11)).

Phosphatase assays

For the pNPP assay, reaction mixtures (60 µl) in 20 mM MES (pH 6.0), various concentrations of p-nitrophenylphosphate (pNPP) and divalent cations were pre-equilibrated at 30 °C for 10 min. Six microlitres of enzyme stock (0.25 µg µl⁻¹ Eya3(ED) and Eya2(ED), 2.5 µg µl⁻¹ Eya(ED) and Eya1(ED), 0.05 µg µl⁻¹ *A. thaliana* EYA, and 0.01 µg µl⁻¹ PTP1B in 20 mM Tris (pH 8.0), 150 mM NaCl) was added and incubated for 20 min. Reactions were quenched with 100 µl 0.5 M EDTA, pH 10.0. Release of p-nitrophenol was determined by measuring A₄₁₀ and extrapolating the values to a standard curve of p-nitrophenol. The data were fitted directly into the Michaelis–Menten equation using SIGMAPLOT (SPSS Science).

For peptide assays, reaction mixtures (60 µl) with different concentrations of phosphopeptides (50 µM–1500 µM), phosphoamino acids (0.1 mM–5 mM), dNTPs or phospholipid vesicles were preheated at 30 °C for 10 min and reactions started by adding enzyme. Reactions were quenched after 20 min with 50 µl malachite green reagent (Promega). Released phosphate was determined by measuring A₆₅₀ and extrapolating the values to a phosphate standard curve.

Protein tyrosine phosphatase blot-overlay assay

293 cells in DMEM containing 10% fetal bovine serum were grown to 80% confluence in six-well culture dishes. A concentration of 0.1 mM pervanadate (1:1 mixture of 100 mM sodium orthovanadate and 30% H₂O₂) was added to the cell culture. After 15 min the cell lysates were electrophoresed on 10% SDS–polyacrylamide gel electrophoresis (PAGE) gels, and transferred to PVDF membrane. The membrane was incubated in phosphatase assay buffer (20 mM MES pH 6.0, 2 mM MgCl₂, 0.5% non-fat dry milk) with proteins and additives as indicated in Fig. 2a, at 30 °C for 1 h. The PVDF strips were visualized by blotting with anti-phosphotyrosine antibody.

Drosophila genetics

UAS-*eya* constructs were generated using full-length *eya* complementary DNA cloned into the *EcoRI* and *BglII* sites of pUAST. Transgenic animals were generated according to standard protocols^{27,28}. All crosses with the UAS-*eya*^{WT} transgene were carried out at 22 °C whereas all crosses with UAS-*eya*^{D493N} transgenes were carried out at 25 °C. This ensured that expression levels for mutant transgenes were equal to or higher than wild type

(Fig. 4f–j). For Fig. 4a–e, w; *eya*² *ey-GAL4* virgin females were crossed to either w; *eya*²; UAS-*eya*^{WT/TM3}, w; *eya*; UAS-*eya*^{D493N/TM3}, or w; *eya*² UAS-*eya*^{D493N/CyO} males, and non-TM3 or non-CyO progeny were scored for eye size. Scanning electron and light microscopy was carried out essentially as described²⁹. Antibody staining was carried out as described³⁰.

Received 28 May; accepted 10 September 2003; doi:10.1038/nature02093.

- Desplan, C. Eye development: governed by a dictator or a junta? *Cell* **91**, 861–864 (1997).
- Xu, P. X. et al. Eya1 is required for the morphogenesis of mammalian thymus, parathyroid and thyroid. *Development* **129**, 3033–3044 (2002).
- Borsani, G. et al. EYA4, a novel vertebrate gene related to *Drosophila* eyes absent. *Hum. Mol. Genet.* **8**, 11–23 (1999).
- Heanue, T. A. et al. Synergistic regulation of vertebrate muscle development by Dach2, Eya2, and Six1, homologs of genes required for *Drosophila* eye formation. *Genes Dev.* **13**, 3231–3243 (1999).
- Xu, P. X., Woo, L., Her, H., Beier, D. R. & Maas, R. L. Mouse Eya homologues of the *Drosophila* eyes absent gene require Pax6 for expression in lens and nasal placode. *Development* **124**, 219–231 (1997).
- Xu, P. X., Cheng, J., Epstein, J. A. & Maas, R. L. Mouse Eya genes are expressed during limb tendon development and encode a transcriptional activation function. *Proc. Natl Acad. Sci. USA* **94**, 11974–11979 (1997).
- Xu, P. X. et al. Eya1-deficient mice lack ears and kidneys and show abnormal apoptosis of organ primordia. *Nature Genet.* **23**, 113–117 (1999).
- Abdelhak, S. et al. Clustering of mutations responsible for branchio-oto-renal (BOR) syndrome in the eyes absent homologous region (eyaHR) of EYA1. *Hum. Mol. Genet.* **6**, 2247–2255 (1997).
- Azuma, N., Hirakiyama, A., Inoue, T., Asaka, A. & Yamada, M. Mutations of a human homologue of the *Drosophila* eyes absent gene (EYA1) detected in patients with congenital cataracts and ocular anterior segment anomalies. *Hum. Mol. Genet.* **9**, 363–366 (2000).
- Wayne, S. et al. Mutations in the transcriptional activator EYA4 cause late-onset deafness at the DFNA10 locus. *Hum. Mol. Genet.* **10**, 195–200 (2001).
- Ohto, H. et al. Cooperation of Six and Eya in activation of their target genes through nuclear translocation of Eya. *Mol. Cell. Biol.* **19**, 6815–6824 (1999).
- Charbonneau, H. & Tonks, N. K. 1002 protein phosphatases? *Annu. Rev. Cell Biol.* **8**, 463–493 (1992).
- Pignoni, F. et al. The eye-specification proteins So and Eya form a complex and regulate multiple steps in *Drosophila* eye development. *Cell* **91**, 881–891 (1997).
- Chen, R., Amoui, M., Zhang, Z. & Mardon, G. Dachshund and eyes absent proteins form a complex and function synergistically to induce ectopic eye development in *Drosophila*. *Cell* **91**, 893–903 (1997).
- Fan, X. et al. The alpha subunits of Gz and Gi interact with the eyes absent transcription cofactor Eya2, preventing its interaction with the six class of homeodomain-containing proteins. *J. Biol. Chem.* **275**, 32129–32134 (2000).
- Aravind, L., Galperin, M. Y. & Koonin, E. V. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **23**, 127–129 (1998).
- Collet, J. F., Stroobant, V., Pirard, M., Delpierre, G. & Van Schaftingen, E. A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif. *J. Biol. Chem.* **273**, 14107–14112 (1998).
- Collet, J. F., Stroobant, V. & Van Schaftingen, E. Mechanistic studies of phosphoserine phosphatase, an enzyme related to P-type ATPases. *J. Biol. Chem.* **274**, 33985–33990 (1999).
- Ridder, I. S. & Dijkstra, B. W. Identification of the Mg²⁺-binding site in the P-type ATPase and phosphatase members of the HAD (haloacid dehalogenase) superfamily by structural similarity to the response regulator protein CheY. *Biochem. J.* **339**, 223–226 (1999).
- Morais, M. C. et al. The crystal structure of *Bacillus cereus* phosphonoacetaldehyde hydrolase: insight into catalysis of phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily. *Biochemistry* **39**, 10385–10396 (2000).
- Ridder, I. S., Rozeboom, H. J., Kalk, K. H., Janssen, D. B. & Dijkstra, B. W. Three-dimensional structure of L-2-haloacid dehalogenase from *Xanthobacter autotrophicus* GJ10 complexed with the substrate-analogue formate. *J. Biol. Chem.* **272**, 33015–33022 (1997).
- Zhang, Z. Y. et al. Substrate specificity of the protein tyrosine phosphatases. *Proc. Natl Acad. Sci. USA* **90**, 4446–4450 (1993).
- Daum, G. et al. A general peptide substrate for protein tyrosine phosphatases. *Anal. Biochem.* **211**, 50–54 (1993).
- Denu, J. M. et al. The purification and characterization of a human dual-specific protein tyrosine phosphatase. *J. Biol. Chem.* **270**, 3796–3803 (1995).
- Bonini, N. M., Leiserson, W. M. & Benzer, S. The eyes absent gene: genetic control of cell survival and differentiation in the developing *Drosophila* eye. *Cell* **72**, 379–395 (1993).
- Bui, Q. T., Zimmerman, J. E., Liu, H. & Bonini, N. M. Molecular analysis of *Drosophila* eyes absent mutants reveals features of the conserved Eya domain. *Genetics* **155**, 709–720 (2000).
- Rubin, G. M. & Spradling, A. C. Genetic transformation of *Drosophila* with transposable element vectors. *Science* **218**, 348–353 (1982).
- Spradling, A. C. & Rubin, G. M. Transposition of cloned P elements into *Drosophila* germ line chromosomes. *Science* **218**, 341–347 (1982).
- Kango-Singh, M. et al. Shar-pei mediates cell proliferation arrest during imaginal disc growth in *Drosophila*. *Development* **129**, 5719–5730 (2002).
- Pappu, K. S. et al. Mechanism of hedgehog signaling during *Drosophila* eye development. *Development* **130**, 3053–3062 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Lang and J. Schlessinger for discussions and comments on the manuscript. Mouse *Eya1* cDNA was provided by T. Glaser and *Drosophila* *eya* cDNA was provided by N. Bonini.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.S.H. (rashmi.hegde@cchmc.org).

The transcription factor Eyes absent is a protein tyrosine phosphatase

Tina L. Tootle^{1,2*}, Serena J. Silver^{1,2*}, Erin L. Davies^{1,*†}, Victoria Newman¹, Robert R. Latek¹, Ishara A. Mills^{1,2}, Jeremy D. Selengut³, Beth E. W. Parlikar^{1,2} & Ilaria Rebay^{1,2}

¹Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA

²Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA

³The Institute for Genomic Research, Rockville, Maryland 20850, USA

* These authors contributed equally to this work

† Present address: Department of Developmental Biology, Stanford University School of Medicine, Stanford, California 94305, USA

Post-translational modifications provide sensitive and flexible mechanisms to dynamically modulate protein function in response to specific signalling inputs¹. In the case of transcription factors, changes in phosphorylation state can influence protein stability, conformation, subcellular localization, cofactor interactions, transactivation potential and transcriptional output¹. Here we show that the evolutionarily conserved transcription factor Eyes absent (Eya)^{2,3} belongs to the phosphatase subgroup of the haloacid dehalogenase (HAD) superfamily^{4,5}, and propose a function for it as a non-thiol-based protein tyrosine phosphatase. Experiments performed in cultured *Drosophila* cells and *in vitro* indicate that Eyes absent has intrinsic protein tyrosine phosphatase activity and can autocatalytically dephosphorylate itself. Confirming the biological significance of this function, mutations that disrupt the phosphatase active site severely compromise the ability of Eyes absent to promote eye specification and development in *Drosophila*. Given the functional importance of phosphorylation-dependent modulation of transcription factor activity, this evidence for a nuclear transcriptional coactivator with intrinsic phosphatase activity suggests an unanticipated method of fine-tuning transcriptional regulation.

The transcriptional coactivator Eyes absent (Eya) is a member of an evolutionarily conserved set of nuclear transcription factors and cofactors collectively termed the retinal determination (RD) gene network^{2,3,6}. Although RD network members are perhaps best known for their roles in eye specification, redeployment of these genes, either individually or as a network, contributes to a diverse array of essential developmental processes in all metazoans^{2,3}. Eya family members are defined by a conserved ~275-amino-acid motif, referred to as the Eya domain (ED), that has been shown to bind two other RD members, Sine oculis (So) and Dachshund (Dac)^{6–8}. Together, Eya and So form a potent transcriptional activator⁹; the mechanistic implications of Eya–Dac interactions are less clear^{10,11}. Emphasizing the functional conservation among Eya homologues, mammalian Eya transgenes can rescue the ‘eyeless’ phenotype of *Drosophila* eya mutations^{6,12}.

We have examined a function for the carboxy-terminal ED of Eya that was suggested by protein motif searches and structural modelling studies. These investigations place Eya within the phosphatase subgroup of the haloacid dehalogenase (HAD) superfamily (Fig. 1a and Supplementary Fig. S1a). HAD family members, found in organisms ranging from bacteria and archaea to humans, constitute a diverse collection of enzymes that includes dehalogenases, ATPases, phosphonates, phosphomutases, epoxy hydrolases and a growing number of magnesium-dependent phosphatases^{4,5,13}. Understanding of the *in vivo* function of HAD family phosphatases remains extremely limited, particularly in eukaryotic systems.

X-ray crystallography combined with mutagenesis studies of several HAD family proteins has revealed a conserved α/β -hydrolase

fold that unites three non-contiguous sequence motifs to form the catalytic core of the enzyme^{13–15}. Five conserved residues brought together by this tripartite configuration have been shown to be essential for catalysis^{13–15}. Structural modelling studies predict that the ED forms a HAD α/β -hydrolase-like fold (Fig. 1b). The catalytic residues are strikingly conserved in the ED of all Eya proteins (Fig. 1a and Supplementary Fig. S1b). In motif 1 (DXDX(T/V)), the invariant first aspartic acid serves as the nucleophile in all HAD family proteins and probably forms a phospho-aspartate intermediate^{16,17}. The second aspartic acid distinguishes the phosphatase/phosphohydrolase subgroup from other branches of the HAD superfamily^{4,5,15} and is strictly conserved in all Eya homologues. Motif 2 contains an essential serine/threonine at the end of the β -strand, and motif 3 contributes at least three required residues, a lysine and two aspartic acids, the second of which has undergone a conservative substitution to glutamic acid in Eya proteins. Requirement for the two acidic residues within motif 3 seems to be strictest within the phosphatase/phosphohydrolase branch of the HAD superfamily⁵. The high degree of conservation of this catalytic quintet (D, S/T, K, D, E) in invertebrate, vertebrate and plant Eya homologues suggests that Eya belongs to the phosphatase subgroup of the HAD superfamily.

To investigate whether Eya has intrinsic phosphatase activity, we tested the ability of recombinant murine glutathione S-transferase (GST)-tagged ED fusion proteins to dephosphorylate the synthetic substrate *p*-nitrophenyl phosphate (pNPP). We found that Eya can function as a phosphatase (Fig. 2), and that mutations altering the presumptive HAD active-site residues severely compromise activity (Fig. 2a; see Supplementary Information for details). Sensitivity to the tyrosine phosphatase inhibitor vanadate, but not to inhibitors of serine/threonine phosphatases (Fig. 2a, b), and a requirement for

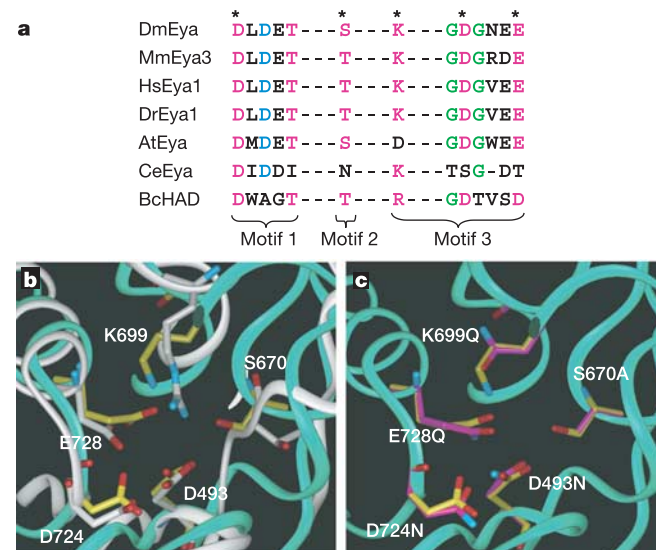


Figure 1 Eya is a member of the phosphatase subgroup of the HAD superfamily. **a**, The non-contiguous sequences comprising HAD motifs 1, 2 and 3. Dm, *D. melanogaster*; Mm, *Mus musculus*; Hs, *Homo sapiens*; Dr, *Danio rerio*; At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Bc, *Bacillus cereus*. Pink residues define the HAD motif; those mutated in this study are marked with an asterisk. Blue residues are most strongly conserved among the phosphatase subgroup of the HAD superfamily; green residues are highly conserved in both ATPases and phosphatases^{13,28}. **b**, Structural modelling studies predict a similar active-site configuration for *Drosophila* Eya (DmEya) and other HAD proteins. The HAD template backbone is identified with a white ribbon and the Eya model backbone is rendered with a cyan ribbon. Key active-site residues are highlighted as sticks, either white for the HAD or yellow for Eya. **c**, Superimposition of mutant DmEya^{HAD} residues on the DmEya model. Alignment of the substitutions (in magenta) and their wild-type counterparts (in yellow) is shown.

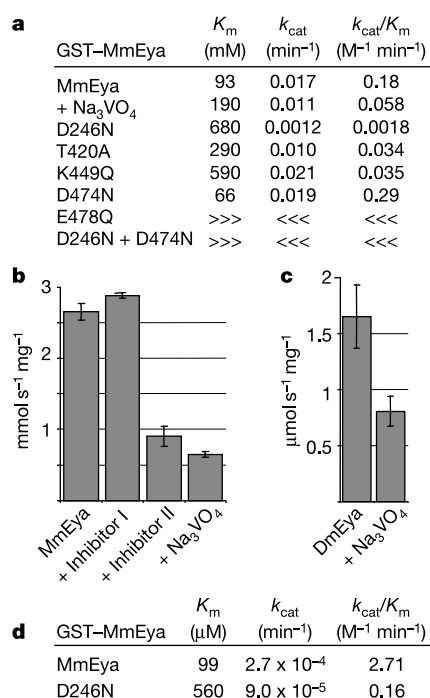


Figure 2 Eya exhibits phosphatase activity *in vitro*. **a**, Kinetic analysis of murine Eya3 GST-ED fusion proteins (GST-MmEya) using the pNPP substrate. The D246N, T420A, K449Q, D474N and E478Q mutations are analogous to the D493N, S670A, K699Q, D724N and E728Q mutations described for DmEya. For those enzymes whose activity was too low to be measured, >>> indicates a K_m significantly higher and an efficiency (k_{cat}/K_m) significantly lower than that measured for D246N. **b**, *In vitro* phosphatase activity of MmEya. **c**, *In vitro* phosphatase activity of DmEya. **d**, Kinetic analysis of GSTMmEya using the tyrosyl phosphorylated peptide substrate I(pY)GEF.

magnesium are consistent with Eya being a HAD family phosphatase¹⁸. We also tested recombinant *Drosophila* Eya in these assays, and although its activity was significantly lower, it was also magnesium dependent and vanadate sensitive (Fig. 2c). The most likely explanation for the weak *in vitro* activity of the fly ED is that we have not identified appropriate conditions for purifying correctly folded, fully active protein, although we cannot rule out the possibility that *Drosophila* Eya, despite retaining all the conserved residues comprising the HAD motif (Fig. 1), may have limited ability to function as a phosphatase. However, the fact that the mouse Eya isoform used in our *in vitro* assays is able to substitute for *Drosophila* Eya *in vivo*¹², together with the results of the *in vivo* experiments described below, leads us to propose that Eya proteins have a conserved phosphatase function.

To investigate whether Eya has protein phosphatase activity, a role that has not been definitively demonstrated for any other HAD family protein¹⁵, we tested several phosphotyrosine- or phosphothreonine-containing synthetic peptide substrates. Eya showed robust activity toward one of the tyrosyl phosphorylated peptides, with a K_m significantly lower than that measured using the pNPP substrate (Fig. 2d). No measurable activity was detected with the phosphothreonine- or other phosphotyrosine-containing peptides (data not shown; see Methods for details). These results show that Eya has protein tyrosine phosphatase (PTP) activity, but they do not rule out the possibility that Eya could dephosphorylate other substrates as well. The fact that not all tyrosyl phosphorylated peptides were hydrolysed suggests that Eya has specific sequence preferences with respect to its putative protein substrates. Because HAD family phosphatases use a catalytic aspartate^{16,17} as the nucleophile rather than the cysteine residue used by standard PTPs¹⁹, these results indicate that Eya is the founding member of

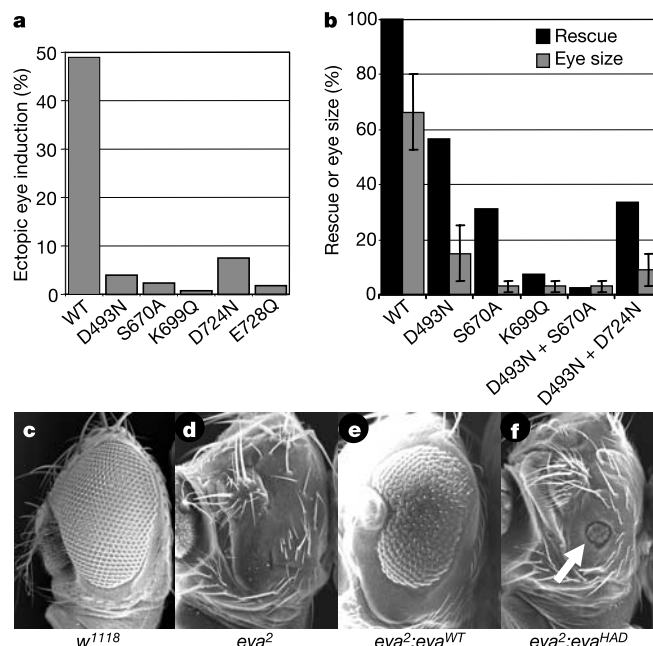


Figure 3 Eya^{HAD} mutants have severely reduced activity relative to Eya^{WT} in ectopic-eye-induction and genetic rescue assays. **a**, The frequency of ectopic eye induction associated with expression of Eya transgenes calculated from multiple independent transgenic lines: Eya^{WT}, 2,465 flies from 8 lines²⁰; Eya^{D493N}, 1,502 flies from 5 lines; Eya^{S670A}, 955 flies from 3 lines; Eya^{K699Q}, 953 flies from 3 lines; Eya^{D724N}, 265 flies from a single line; Eya^{E728Q}, 1,239 flies from 4 lines. **b**, The percentage of eyes from flies of the genotype *eya*²;UAS-*eya/dpp*-GAL4 exhibiting rescue of the *eya*² 'eyeless' phenotype (black bars) and average size of the rescued tissue relative to a wild-type eye (grey bars) are plotted. The following transgenic lines were used: Eya^{WT}, 155 flies from two independent lines; Eya^{D493N}, 124 flies from a single line; Eya^{S670A}, 281 flies from a single line; Eya^{K699Q}, 176 flies from a single line; Eya^{D493N+S670A}, 209 flies from two independent lines; Eya^{D493N+D724N}, 151 flies from two independent lines. **c-f**, Scanning electron micrographs of adult eyes. **c**, *w*¹¹¹⁸. **d**, *eya*². **e**, *eya*²;UAS-*eya*^{WT}/dpp-GAL4. **f**, *eya*²;UAS-*eya*^{HAD}/dpp-GAL4; arrow points to a small patch of rescued eye tissue.

a new class of non-thiol-based PTPs.

We used the genetically tractable *Drosophila* system to investigate the physiological relevance of the putative PTP activity of Eya. Site-directed mutagenesis was used to generate five single and four double mutant combinations of the five HAD active-site residues in *Drosophila* Eya (Fig. 1c); we refer to them collectively as the Eya^{HAD} mutants. These mutants were first transfected into *Drosophila* S2 cultured cells. Immunostaining and western blotting analyses revealed no apparent changes in subcellular localization (data not shown) or expression levels (Supplementary Fig. S2a) of the mutants relative to Eya^{WT}.

Eya, like most RD genes, induces formation of eye tissue outside the normal eye field when ectopically expressed^{2,3,6,20}. Scoring the percentage of flies exhibiting ectopic eye formation therefore provides a sensitive measurement of Eya activity²⁰. To determine whether the HAD active-site mutants compromise the ectopic-eye-induction potential of Eya, we generated transgenic lines carrying full-length Eya^{HAD} mutant expression constructs. All the Eya^{HAD} mutants showed markedly reduced ectopic eye induction relative to Eya^{WT} (Fig. 3a). Protein expression levels from the Eya^{HAD} transgenes were comparable to those from Eya^{WT} lines (Supplementary Fig. S2b), indicating that the reduction in ectopic-eye-inducing potential reflects a change in protein activity rather than reduced expression. Comparable reductions in Eya activity were also observed with Eya^{HAD} transgenes in which two of the five HAD active-site residues were mutated simultaneously (data not shown).

Because the HAD motif active-site mutants compromise the ability

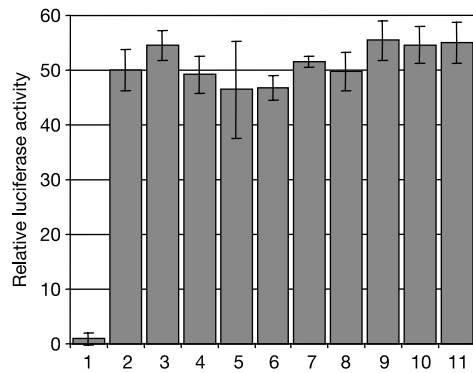


Figure 4 Eya^{HAD} mutations do not disrupt the role of Eya as a transcriptional coactivator in conjunction with So. The *Drosophila* cell-culture-based transcription assays were performed as described⁹. Lanes: 1, Are-luciferase; 2, Eya^{WT}; 3, Eya^{D493N}; 4, Eya^{S670A}; 5, Eya^{K699Q}; 6, Eya^{D724N}; 7, Eya^{E728Q}; 8, Eya^{D493N+S670A}; 9, Eya^{D493N+K699Q}; 10, Eya^{D493N+D724N}; 11, Eya^{D493N+E728Q}. See Supplementary Fig. S3 for further details.

of Eya to induce ectopic eye formation, we asked whether an intact HAD motif is required for normal Eya function during eye development. We compared the ability of Eya^{WT} versus Eya^{HAD} transgenes to complement the eye-specific loss-of-function *eya*² allele. Homozygous *eya*² mutant flies exhibit a completely penetrant 'eyeless' phenotype, in which the entire eye is missing (Fig. 3c, d). For these experiments we define rescue as the ability of a given transgene to produce recognizable eye tissue within the normal eye field of an adult fly. We also can compare the extent of rescue by estimating the size of the rescued eye tissue relative to that of a wild-type eye.

Expression of Eya^{WT} transgenes rescues the *eya*² 'eyeless' phenotype with complete penetrance in both eyes of each individual fly (Fig. 3b, e; data not shown). In striking contrast, all Eya^{HAD} mutant transgenes exhibited a significantly reduced frequency and extent of rescue, with rescue usually occurring in only one of the two eye fields of an individual (Fig. 3b, f). For all Eya^{HAD} transgenes tested, even in cases where rescue efficiency was only twofold to threefold lower than that of Eya^{WT}, the size of the rescued eye tissue was always significantly (fivefold to tenfold) reduced relative to that obtained with Eya^{WT} lines (Fig. 3b, e, f). Western blot analyses of eye imaginal discs ruled out the possibility that reduced protein expression might be responsible for this result (Supplementary Fig. S2c). In combination with the ectopic-eye-induction assay data, the results of these rescue experiments argue strongly that the activity of Eya as a putative HAD family phosphatase is required to promote normal eye development in *Drosophila*.

Because the region of the ED that binds to the RD gene network protein So^{7,12} partially overlaps with motif 1 of the HAD domain (Supplementary Fig. S1b), we checked whether the Eya^{HAD} missense mutations compromise the ability of Eya to interact productively with So. Eya and So interact to form a potent transcriptional activator required for eye specification, in which So contributes the DNA-binding domain and Eya provides the transactivation potential^{7,9,11}. Using a transcription assay in *Drosophila* S2 cultured cells⁹, we found that the ability of Eya^{HAD} mutant proteins to synergize with So to activate transcription of a reporter gene is comparable to that of Eya^{WT} (Fig. 4 and Supplementary Fig. S3). Although we cannot rule out the formal possibility that *in vivo* the Eya^{HAD} mutations disrupt interactions with other proteins rather than block phosphatase activity, the finding that disruption of the HAD motif active site does not abrogate the ability of Eya to function as a transcriptional coactivator in conjunction with So leads us to propose that Eya proteins have two essential functions: a previously described role as a transcription factor and a role as a protein tyrosine phosphatase.

To investigate the intrinsic PTP activity of Eya on physiologically

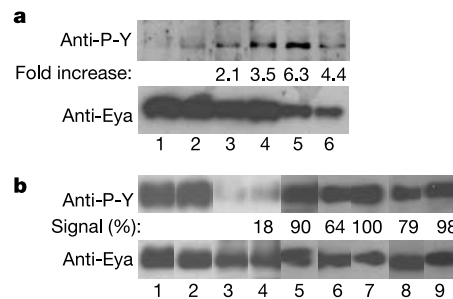


Figure 5 Eya has protein tyrosine phosphatase activity. Top panels show immunoblots probed with anti-phosphotyrosine antibody (anti-P-Y); bottom panels show immunoblots of the same samples probed with anti-Flag antibody to detect Eya (anti-Eya). **a**, Eya^{HAD} proteins exhibit higher levels of tyrosine phosphorylation than Eya^{WT}. Lanes: 1 and 2, independent transfections of Eya^{WT}; 3, Eya^{D493N+S670A}; 4, Eya^{D493N+K699Q}; 5, Eya^{D493N+D724N}; 6, Eya^{D493N+E728Q}. Fold increase calculated as signal for the Eya^{HAD} mutant P-Y level relative to the average of the P-Y signal in the Eya^{WT} lanes, and corrected relative to the strength of the anti-EYA signal. **b**, Dephosphorylation of DmEya^{D493N+D724N} by recombinant GST-ED. Full-length tyrosine phosphorylated DmEya^{D493N+D724N} (all lanes) was immunoprecipitated and incubated with recombinant murine GST-ED, either wild type (WT) or HAD mutant variants. Lanes: 1 and 2, control; 3 and 4, Eya^{WT}; 5, Eya^{D246N}; 6, Eya^{T420A}; 7, Eya^{K449Q}; 8, Eya^{D474N}; 9, Eya^{E478Q}. The percentage of anti-P-Y signal for Eya^{D493N+D724N} relative to controls and corrected for relative protein levels is indicated. Numbers shown are an average from two independent experiments for each GST-ED tested; results from only one of the two experiments are shown for the GST-ED HAD mutants.

relevant substrate candidates, we exploited the finding that Eya can be tyrosine phosphorylated in *Drosophila* S2 cells (Fig. 5a; see Supplementary Information for discussion). This allowed us to affinity purify full-length Eya from these cells and to use it as a protein substrate in an *in vitro* phosphatase reaction. Because the phosphotyrosine signal associated with the Eya^{HAD} mutant proteins was consistently elevated relative to Eya^{WT} (Fig. 5a; see Supplementary Information for discussion), Eya^{HAD} protein was used as the substrate. We found that incubation of Eya^{HAD} protein with recombinant murine GST-ED protein strongly reduced the phosphotyrosine signal (Fig. 5b). HAD active-site mutants that exhibit impaired activity both *in vitro* and *in vivo* (Figs 2 and 3) also have severely reduced activity in this assay (Fig. 5b). These results show that Eya can function as a PTP with a full-length endogenous protein substrate and that such activity depends on an intact HAD motif. Although we do not yet understand the physiological relevance of tyrosine phosphorylation and dephosphorylation of Eya, these results (Fig. 5), together with the previous demonstration that Eya is able to self-associate⁹, suggest that Eya may autocatalytically dephosphorylate itself.

In conclusion, we propose that Eya is the founding member of a novel class of non-thiol-based PTPs and is, to our knowledge, the first example of a transcription factor with intrinsic phosphatase activity. Further work will be required to understand how tyrosine phosphorylation and dephosphorylation regulate Eya function *in vivo*, and what substrates, potentially including Eya itself, may be regulated by its PTP activity. Elucidation of the biochemical regulatory mechanisms that coordinate the dual functions of Eya as transactivator and phosphatase during eye specification will provide insights into the function of the RD gene network, and more generally establish a new paradigm for transcriptional regulatory strategies. Although preliminary analyses have not identified other HAD-motif-containing proteins that are annotated as transcriptional regulators (R.R.L. and I.R., unpublished observation), it seems likely that dual-function mechanisms analogous to that which we propose for Eya will prove to be a general strategy for fine-tuning transcriptional output, particularly in highly regulated developmental systems. □

Methods

Bioinformatics

For a description of the computational analyses, see Supplementary Information.

Phosphatase assays

Phosphatase assays were performed using GST–ED fusion proteins (purified as described in Supplementary Information). Assays to measure the enzyme kinetics using the synthetic substrate *p*-nitrophenyl phosphate (Sigma) were done in triplicate with six substrate concentrations over six time points. Reactions (80 µl) were performed at 30 °C in 200 mM PIPES, pH 7.0, 5 mM EDTA and 10 mM MgCl₂, and were quenched by the addition of 40 µl of 10 M NaOH. PNP anion was detected at 405 nm (extinction coefficient $\epsilon = 1.78 \times 10^4 \text{ l mol}^{-1} \text{ cm}^{-1}$) using a Tecan GENios plate reader. Reactions were normalized to buffer-alone controls and the results analysed by Lineweaver–Burk plot using Microsoft Excel. Synthetic peptide substrates tested were: I(pY)GEF and TSTGPE(pY)EPGENL (Calbiochem); END(pY)INASL, DADE(pY)LIPQQG and RRA(pT)VA (Promega). Reactions (50 µl) were performed at 25 °C in 200 mM HEPES, pH 7.0, 10 mM MgCl₂ and 5 mM EDTA, and were quenched with 50 µl of molybdate dye solution (Promega). Malachite green–ammonium molybdate phosphate complex was detected at 595 nm and converted to moles of free phosphate using a phosphate concentration standard curve. Assays with I(pY)GEF were carried out at five substrate concentrations over five time points and the results were analysed as described for pNPP. Phosphatase inhibitor cocktail sets I and II (Calbiochem; see Supplementary Information for details) were used at a 1:50 dilution in pNPP phosphatase assays. Sodium orthovanadate was used at 4 mM final concentration in pNPP phosphatase assays.

Amino-terminally Flag-epitope-tagged Eya constructs were subcloned into the copper inducible metallothionein promoter vector. Each construct (5 µg of DNA) was transfected into S2 cells as previously described²¹. Following published protocols^{22–26}, cells were treated with 100 µM NaVO₃ and 200 µM H₂O₂ for 15 min before lysis in 100 mM NaCl, 50 mM Tris–HCl, pH 7.5, 2 mM EDTA, 2 mM EGTA, 1% NP-40, 1 mM Na₃VO₄ and one mini-complete protease inhibitor tablet (Roche) per 10 ml. All subsequent solutions include 1 mM Na₃VO₄. Clarified lysates were incubated with 25 µl of anti-Flag M2 agarose affinity gel (Sigma) for 1.5 h at 4 °C. Beads were washed twice in lysis buffer and twice in 10 mM Tris–HCl, pH 7.5, and 150 mM NaCl, resuspended in 30 µl of ×2 SDS sample buffer and boiled, and 10 µl were loaded per lane. Western blots were performed as described²⁷ except that blocking and antibody incubations were performed in 1% casein according to Hammarsten (EM Science). Antibodies: guinea-pig anti-Eya 1:16,000; rabbit anti-phosphotyrosine 1:400 (0.21 mg^{−1} ml^{−1}; Upstate); HRP-conjugated goat anti-guinea-pig and anti-rabbit 1:5,000 (Jackson ImmunoResearch). Determination of the fold increase in phosphotyrosine signal relative to Eya protein amounts was performed using NIH Image software; samples analysed in this way were run together on the same gel.

To obtain sufficient tyrosine phosphorylated *Drosophila* Eya to use as a substrate in the *in vitro* phosphatase assay, a stable cell line expressing Flag-tagged Eya^{D493N+D724N} was generated. Cells (500 µl) were immunoprecipitated as described above, except that 1 mM Na₃VO₄ was omitted from the wash buffer. The washed immunoprecipitates were incubated in phosphatase assay reaction buffer (as described above but without pNPP), either with GST agarose or with 100 µg GST–ED proteins for 1 h at 30 °C, processed for western blotting and analysed as described above.

Molecular biology and genetics

Site-directed mutagenesis, subcloning, generation of transgenics, crosses, ectopic eye scoring, calculation of per cent ectopic eye induction and scanning electron microscopy were performed as previously described^{20,21}. Fly crosses were performed at 25 °C with the exception of the genetic rescue assays, which were performed at 20 °C.

Received 3 June; accepted 22 September 2003; doi:10.1038/nature02097.

- Hunter, T. Signaling—2000 and beyond. *Cell* **100**, 113–127 (2000).
- Treisman, J. E. A conserved blueprint for the eye? *Bioessays* **21**, 843–850 (1999).
- Wawersik, S. & Maas, R. L. Vertebrate eye development as modeled in *Drosophila*. *Hum. Mol. Genet.* **9**, 917–925 (2000).
- Collet, J. F., van Schaftingen, E. & Stroobant, V. A new family of phosphotransferases related to P-type ATPases. *Trends Biochem. Sci.* **23**, 284 (1998).
- Thaller, M. C., Schippa, S. & Rossolini, G. M. Conserved sequence motifs among bacterial, eukaryotic, and archaeal phosphatases that define a new phosphohydrolase superfamily. *Protein Sci.* **7**, 1647–1652 (1998).
- Bonini, N. M., Bui, Q. T., Gray-Board, G. L. & Warrick, J. M. The *Drosophila* eyes absent gene directs ectopic eye formation in a pathway conserved between flies and vertebrates. *Development* **124**, 4819–4826 (1997).
- Pignoni, F. *et al.* The eye-specification proteins So and Eya form a complex and regulate multiple steps in *Drosophila* eye development. *Cell* **91**, 881–891 (1997).
- Chen, R., Amoui, M., Zhang, Z. & Mardon, G. Dachshund and Eyes absent proteins form a complex and function synergistically to induce ectopic eye development in *Drosophila*. *Cell* **91**, 893–903 (1997).
- Silver, S. J., Davies, E. L., Doyon, L. & Rebay, I. A functional dissection of Eyes absent reveals new modes of regulation within the retinal determination gene network. *Mol. Cell. Biol.* **23**, 5989–5999 (2003).
- Kim, S. S. *et al.* Structure of the retinal determination protein Dachshund reveals a DNA binding motif. *Structure (Camb.)* **10**, 787–795 (2002).
- Ikedo, K., Watanabe, Y., Ohto, H. & Kawakami, K. Molecular interaction and synergistic activation of a promoter by Six, Eya, and Dach proteins mediated through CREB binding protein. *Mol. Cell. Biol.* **22**, 6759–6766 (2002).
- Bui, Q. T., Zimmerman, J. E., Liu, H. & Bonini, N. M. Molecular analysis of *Drosophila* eyes absent mutants reveals features of the conserved Eya domain. *Genetics* **155**, 709–720 (2000).
- Collet, J. F., Stroobant, V. & Van Schaftingen, E. Mechanistic studies of phosphoserine phosphatase, an enzyme related to P-type ATPases. *J. Biol. Chem.* **274**, 33985–33990 (1999).

- Aravind, L., Galperin, M. Y. & Koonin, E. V. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **23**, 127–129 (1998).
- Selengut, J. D. Mdp-1 is a new and distinct member of the haloacid dehalogenase family of aspartate-dependent phosphohydrolases. *Biochemistry* **40**, 12704–12711 (2001).
- Ridder, I. & Dijkstra, B. Identification of the Mg²⁺-binding site in the P-type ATPase and phosphatase members of the HAD (haloacid dehalogenase) superfamily by structural similarity to the response regulator protein CheY. *Biochem. J.* **339**, 223–226 (1999).
- Cho, H. *et al.* Beryllium fluoride acts as a phosphate analog in proteins phosphorylated on aspartate: structure of a beryllium fluoride complex with phosphoserine phosphatase. *Proc. Natl Acad. Sci. USA* **98**, 8525–8530 (2001).
- Selengut, J. D. & Levine, R. L. MDP-1: A novel eukaryotic magnesium-dependent phosphatase. *Biochemistry* **39**, 8315–8324 (2000).
- Andersen, J. N. *et al.* Structural and evolutionary relationships among protein tyrosine phosphatase domains. *Mol. Cell. Biol.* **21**, 7117–7136 (2001).
- Hsiao, F. C., Williams, A., Davies, E. L. & Rebay, I. Eyes absent mediates cross-talk between retinal determination genes and the receptor tyrosine kinase signaling pathway. *Dev. Cell* **1**, 51–61 (2001).
- Tootle, T. L., Lee, P. S. & Rebay, I. CRM1-mediated nuclear export and regulated activity of the receptor tyrosine kinase antagonist YAN require specific interactions with MAE. *Development* **130**, 845–857 (2003).
- Cohen, J., Altaratz, H., Zick, Y., Klingmüller, U. & Neumann, D. Phosphorylation of erythropoietin receptors in the endoplasmic reticulum by pervanadate-mediated inhibition of tyrosine phosphatases. *Biochem. J.* **327**, 391–397 (1997).
- Huyer, G. *et al.* Mechanism of inhibition of protein-tyrosine phosphatases by vanadate and pervanadate. *J. Biol. Chem.* **272**, 843–851 (1997).
- Imbert, V. *et al.* Induction of tyrosine phosphorylation and T-cell activation by vanadate peroxide, an inhibitor of protein tyrosine phosphatases. *Biochem. J.* **297**, 163–173 (1994).
- Ruff, S. J., Chen, K. & Cohen, S. Peroxovanadate induces tyrosine phosphorylation of multiple signaling proteins in mouse liver and kidney. *J. Biol. Chem.* **272**, 1263–1267 (1997).
- Scanga, S. E. *et al.* The conserved PI3K/PTEN/Akt signaling pathway regulates both cell size and survival in *Drosophila*. *Oncogene* **19**, 3971–3977 (2000).
- O'Neill, E. M., Rebay, I., Tjian, R. & Rubin, G. M. The activities of two Ets-related transcription factors required for *Drosophila* eye development are modulated by the Ras/MAPK pathway. *Cell* **78**, 137–147 (1994).
- Aravind, L. & Koonin, E. V. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **23**, 469–472 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Voas, D. Doroquez, L. Doyon and B. Chaffee for technical assistance, J. Flynn, S. Neher and S. Flaugh for advice, reagents and technical help, D. Maas for murine Eya complementary DNAs, the Rebay lab for advice and discussions, F. Pignoni for the *eya*² stocks used in the rescue assay, and R. Hegde for sharing unpublished information. We are grateful for advice, comments and encouragement from R. Fehon, G. Fink, T. Orr-Weaver, S. Shenolikar, F. Solomon and J. York. T.L.T., S.J.S., I.A.M. and B.E.W.P. are supported by the Ludwig Foundation, the Howard Hughes Medical Institute, the Packard Foundation and the National Science Foundation, respectively. I.R. is supported by the National Eye Institute and the Burroughs Wellcome Fund.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to I.R. (rebay@wi.mit.edu).

FKF1 is essential for photoperiodic-specific light signalling in *Arabidopsis*

Takato Imaizumi¹, Hien G. Tran¹, Trevor E. Swartz², Winslow R. Briggs² & Steve A. Kay¹

¹Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

²Department of Plant Biology, Carnegie Institution of Washington, 260 Panama Street, Stanford, California 94305, USA

Adaptation to seasonal change is a crucial component of an organism's survival strategy. To monitor seasonal variation, organisms have developed the capacity to measure day length (photoperiodism). Day-length assessment involves the photoperiodic control of flowering in *Arabidopsis thaliana*, whereby the coincidence of light and high expression of CONSTANS (CO)

induces the expression of *FLOWERING LOCUS T* (*FT*), leading to flowering in long-day conditions¹. Although controlling *CO* expression is clearly a key step in day-length discrimination, the mechanism that generates day-length-dependent *CO* expression remains unknown. Here we show that the clock-controlled FLAVIN-BINDING, KELCH REPEAT, F-BOX (*FKF1*)² protein has an essential role in generating the diurnal *CO* peak and that this function is dependent on light. We show that a recombinant *FKF1* LIGHT, OXYGEN OR VOLTAGE (LOV)³ domain binds the chromophore flavin mononucleotide and undergoes light-induced photochemistry, indicating that *FKF1* may function as a photoperiodic blue-light receptor. It is likely that the circadian control of *FKF1* expression and the light regulation of *FKF1* function coincide to control the daytime *CO* waveform precisely, which in turn is crucial for day-length discrimination by *Arabidopsis*.

Arabidopsis plants flower more rapidly in long days (LD) than in short days (SD), and *fkf1* was isolated as a late-flowering mutant in LD². Several mutants are defective in photoperiod-dependent flowering as a consequence of having an impaired circadian clock⁴. Because *FKF1* shares sequence similarity with the *ZEITLUPE* (*ZTL*)⁵ and *LOV KELCH REPEAT PROTEIN 2* (*LKP2*)⁶ genes that have clock function, we examined *fkf1* mutants for circadian defects. We used two *fkf1* alleles that have comparable flowering phenotypes: *fkf1*, which contains a deletion of 65 kilobases (kb); and *fkf1-2*, which contains a T-DNA insertion in the *FKF1* coding region (Supplementary Figs 1 and 2). The phase and the amplitude of two circadian reporters (gene fusions of the luciferase (*luc*) promoter and *CHLOROPHYLL A/B BINDING PROTEIN 2* (*CAB2*) or *COLD CIRCADIAN RNA BINDING PROTEIN 2*; *CCR2*)^{7,8} in the *fkf1* mutants were similar to those in wild-type plants (Fig. 1a).

We also examined the effect of constitutively expressing *FKF1* and *LKP2* (as tandem affinity purification (TAP) fusions)⁹ on clock function. The *cab2::luc* rhythms in 35S::*FKF1*-TAP lines were very similar to those in wild-type lines, whereas ectopic expression of

comparable amounts of *LKP2*-TAP resulted in the arrhythmic expression of *cab2::luc*, as reported previously (ref. 6 and Fig. 1b, c). Although we cannot eliminate the possibility that *FKF1* may have clock functions that were not detected by our reporters, these results indicate that *FKF1* does not function in the clock or clocks controlling the rhythmic expression of *CAB2* and *CCR2*.

Expression of *FKF1* is controlled by the circadian clock, with transcripts peaking in the evening². At least five evening element (EE)-like sequences¹⁰ lie in the first 500 base pairs (bp) of the *FKF1* promoter (data not shown). This implies that clock control of *FKF1* may be mediated by the binding of *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) and *LATE ELONGATED HYPOCOTYL* (*LHY*) to the EEs in the morning¹¹. We therefore examined the expression of *FKF1* in a *cca1 lhy* double mutant. As observed for other genes containing EE elements^{12,13}, the peak amounts of *FKF1* shifted to the morning (Fig. 1d). These results indicate that *FKF1* is probably a clock output directly regulated by core clock proteins. As our analyses have indicated that *FKF1* is not likely to be a ubiquitous component of the clock, this suggests that the late-flowering phenotype of *fkf1* mutants is not caused by a circadian defect.

To gain insight into how *FKF1* might regulate flowering time, we first examined its expression in LD and SD. The *fkf1* mutant was transformed with a construct comprising *FKF1*-TAP complementary DNA driven by the wild-type *FKF1* promoter. The construct largely rescued the late-flowering phenotype of *fkf1* (Supplementary Fig. 2), and expression of the *FKF1*-TAP transcript was very similar to that of *FKF1* (data not shown), indicating that *FKF1*-TAP is expressed and functions in a manner similar to the wild-type protein. *FKF1* peaked at zeitgeber time 10 (ZT10) in LD, whereas *FKF1* peaked at ZT7 in SD (Fig. 2a, b, filled circles). In both conditions, *FKF1*-TAP protein (open circles) accumulated to similar amounts, but peak expression occurred during the day in LD and at night in SD (Fig. 2a, b). This difference in expression suggests that light may have a role in the function of *FKF1*.

Because the timing of *CO* and *FT* expression is crucial in photoperiod-dependent flowering, we investigated whether *FKF1* has an integral role in regulating these genes. In wild-type plants grown in LD, *CO* shows a biphasic pattern of expression consisting of a daytime afternoon peak and a night-time peak (ref. 14 and Fig. 2c). In *fkf1* plants grown in LD, however, *CO* expression was altered such that the daytime peak was absent but the night-time peak remained (Fig. 2c). In SD, *CO* peaked at night-time in both wild-type and *fkf1* plants, although maximal accumulation of *CO* occurred about 3 h later in *fkf1* mutants than in wild-type plants (Fig. 2d).

Analysis of *FT* expression showed that *FT* was not induced in *fkf1* mutants in either LD or SD (Fig. 2e, f). It is likely that the lack of *FT* induction in *fkf1* is due to the low expression of *CO* in the daytime. Indeed, when *CO* was overexpressed in *fkf1*, the plants flowered early (Supplementary Fig. 2). These results show that *CO* functions downstream of *FKF1*. In addition, because there was a more noticeable effect on *CO* expression in the *fkf1* mutant in LD than in SD, and because a greater amount of *FKF1* protein is present during the day in LD than in SD, these results raise the possibility that a combination of light and *FKF1* expression is required for the daytime induction of *CO*. Alternatively, it is possible that *FKF1* functions in concert with a circadian-regulated factor whose expression coincides with *FKF1* specifically in LD to induce *CO*.

To distinguish between these two possibilities, we designed an experiment that would alter light conditions without significantly disturbing the circadian phase (see Supplementary Fig. 3 for full details). Plants were entrained in LD or SD for 9 d, and then on day 10 the day-length conditions were switched from LD to SD or from SD to LD. We reasoned that if *CO* expression is controlled simply by the clock, then light changes on day 10 should not affect it. Note that *FKF1* transcripts begin to accumulate before ZT8 in both day-length conditions (Fig. 2a, b); however, *FKF1* protein is expressed in

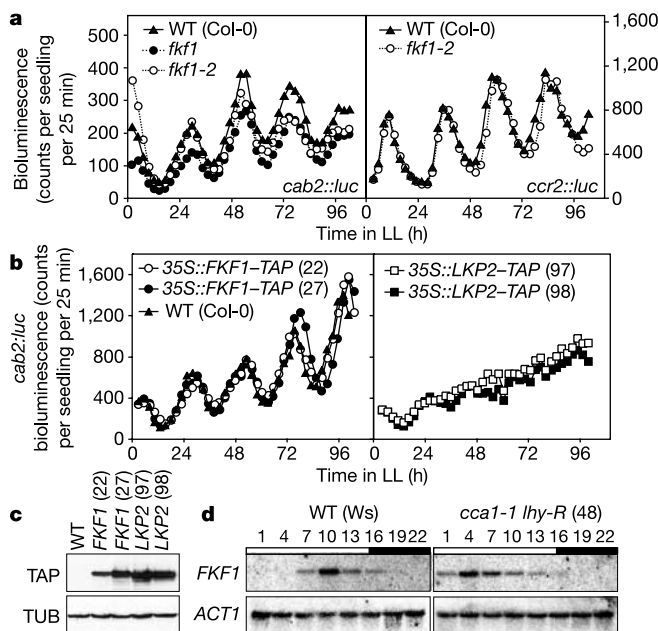


Figure 1 *FKF1* is a clock output. **a**, *cab2::luc* and *ccr2::luc* rhythms (means of 48 traces) in the *fkf1* mutants. **b**, *cab2::luc* rhythms in lines constitutively expressing *FKF1*-TAP and *LKP2*-TAP (means of 24 traces). **c**, Western blot analysis of *FKF1*-TAP and *LKP2*-TAP, with tubulin (TUB) as a control. **d**, Northern blot analysis of *FKF1* expression in a *cca1 lhy* mutant (with *Actin 1* (*ACT1*) as a control). Wild-type Wassilewskija (Ws) ecotype and *cca1-1 lhy-R* (48) were grown in LD.

the dark in LD-entrained plants on day 10, and in the light in SD-entrained plants on day 10 (Fig. 2g, h).

In plants switched from LD to SD, the *CO* waveform in wild-type and *fkf1-2* plants was comparable to that observed in SD (compare Fig. 2d to Fig. 2i). By contrast, in plants switched from SD to LD

(where the light period is extended), the *CO* waveform in wild-type plants clearly showed a biphasic pattern (Fig. 2j). In addition, the daytime *CO* peak was observed at ZT10, which is the same ZT time at which FKF1 protein peaks in SD-entrained plants (Fig. 2b, h, j). In *fkf1-2* plants switched from SD to LD, light suppressed *CO* expression (Fig. 2j). In both wild-type and *fkf1* plants, induction of *FT* expression was observed only when *CO* quantities were relatively high under light (compare Fig. 2i to j and k to l), further supporting the hypothesis that an increase in *CO* and light together induce *FT*. These results indicate that in the absence of FKF1 light suppresses *CO* expression, and that the daytime peak of *CO* may occur only when light appropriately coincides with FKF1 expression.

If light has the functions proposed above, *CO* expression should be dependent primarily on expression of FKF1 in constant light (LL) conditions. In LL, FKF1 protein cycled robustly as did the *CO* transcript, although *CO* had a night-time peak of reduced amplitude (ref. 14 and Supplementary Fig. 4a, b). In addition, the peak of *CO* expression coincided with large amounts of FKF1 protein, whereas overall amounts of *CO* were clearly reduced in *fkf1-2*. These results support the idea that light and cyclic FKF1 expression generate the oscillation of *CO* expression under LL conditions. Expression of *FT* seemed to depend on the amounts of *CO* (Supplementary Fig. 4c), further supporting the idea that *CO* and light are required to induce *FT*.

Because these findings suggested that FKF1 activity can be regulated by light, and because FKF1 contains a LOV domain, the light-sensing module of phototropin (phot) blue-light photoreceptors^{3,15}, this raised the possibility that FKF1 can directly detect light. We therefore tested whether the FKF1 LOV domain possesses features of a functional light sensor. By using a purified fusion protein, we determined that the FKF1 LOV domain is a flavoprotein that binds to a flavin mononucleotide (FMN) chromophore (Fig. 3a, b). Both the ground-state absorption spectra (Fig. 3a) and the light-induced difference spectra of FKF1 LOV domain matched that of the phot1 LOV2 domain (Fig. 3c, red spectra),

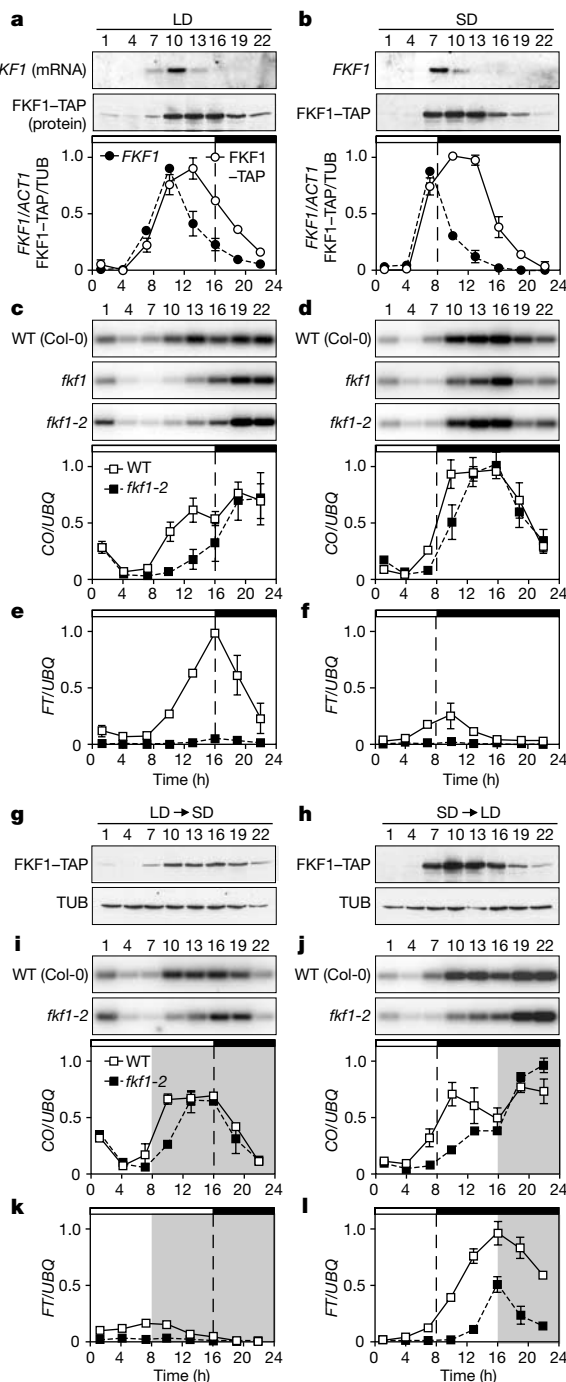


Figure 2 *CO* and *FT* expression is dependent on FKF1 and light. **a–f**, Plants were grown in LD or SD. Shown are *FKF1* messenger RNA and FKF1–TAP expression (**a, b**), and *CO* (**c, d**) and *FT* (**e, f**) mRNA expression in wild-type (WT) and *fkf1* mutant plants in LD and SD. **g–l**, Plants were entrained in either LD or SD, and were shifted from LD to SD on day 10 or from SD to LD on day 10. Shown are FKF1–TAP protein expression (**g, h**), and *CO* (**i, j**), and *FT* (**k, l**) mRNA expression in different day-length conditions. Representative blotting images and means $\pm$ s.e.m. of the quantified data are shown. The entrained conditions are shown by white (light periods) and black (dark periods) bars above traces; grey areas behind the traces represent dark periods on day 10 (**d, e, g, h**).

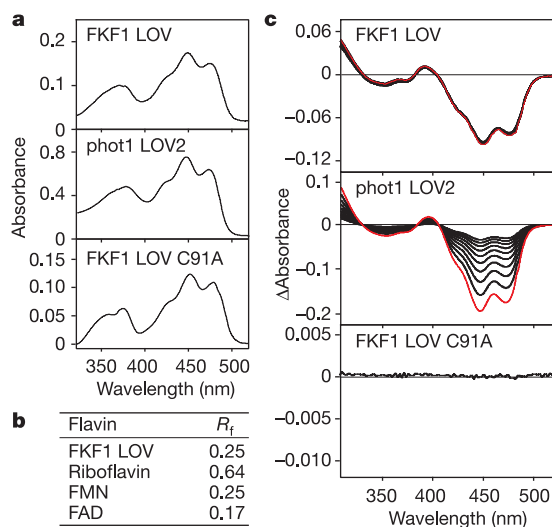


Figure 3 The FKF1 LOV domain possesses light-sensing properties. **a**, Absorption spectra of FKF1 LOV, phot1 LOV2 and FKF1 LOV C91A. **b**, Identification of FMN as the flavin chromophore associated with FKF1 LOV domain by thin layer chromatography. Retardation factor (R_f) values for the FKF1 LOV chromophore and flavin standards are shown. **c**, Light-minus-dark difference spectra of FKF1 LOV, phot1 LOV2 and FKF1 LOV C91A. These spectra show the light-induced absorbance changes (red spectra) and subsequent spectral properties in the dark (black spectra) for FKF1 LOV and phot1 LOV2. After a light flash, the difference spectra were taken every 5 min for FKF1 or every 15 s for phot1 LOV2. The difference spectra of FKF1 LOV C91A did not show any light-induced absorbance changes.

suggesting that they have similar chromophore-binding and light-induced photochemistry. In contrast to phototropin LOV domains, the FKF1 LOV domain showed no appreciable dark recovery (Fig. 3c, black spectra). The absorption spectra of FKF1 had a maximum at 450 nm with vibronic side bands and a smaller peak in the ultraviolet-A region (Fig. 3a). After light irradiation, there was a loss of absorption at about 450 nm and an increase in absorption at around 390 nm, which is shown as a difference spectra in Fig. 3c.

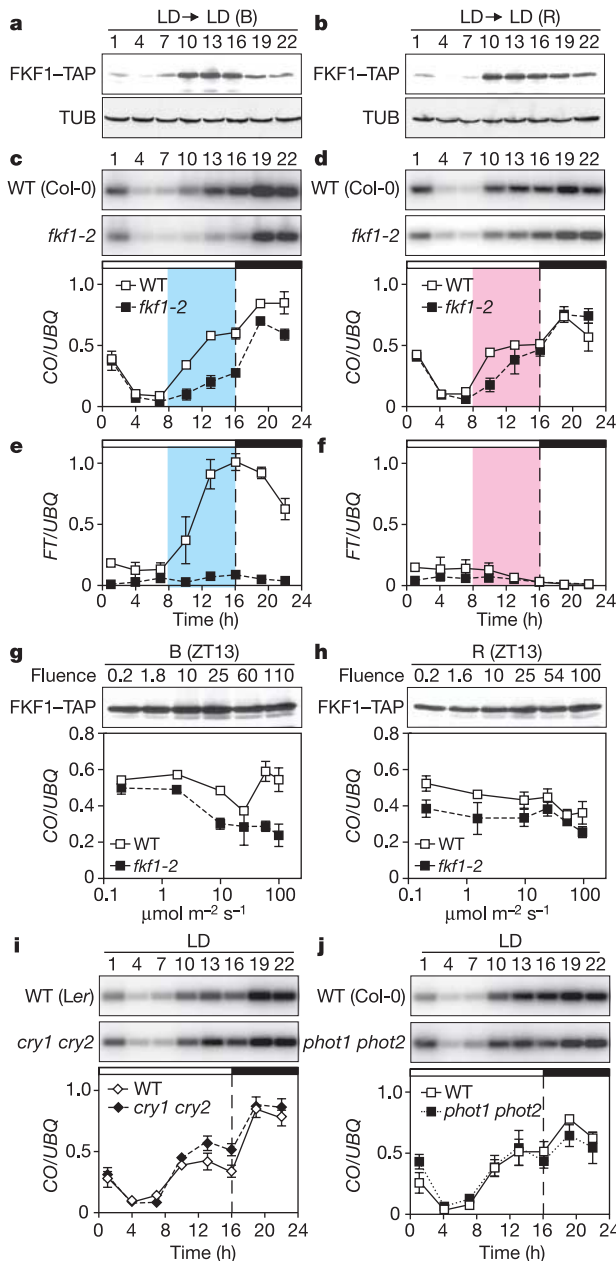


Figure 4 FKF1 may function as a blue-light photoreceptor. **a–f**, Blue and red light effects on the expression of FKF1-TAP protein (**a**, **b**), and *CO* (**c**, **d**) and *FT* (**e**, **f**) mRNA. LD-entrained plants were irradiated with either blue light ($60 \mu\text{mol m}^{-2} \text{s}^{-1}$; **a**, **c**, **e**) or red light ($50 \mu\text{mol m}^{-2} \text{s}^{-1}$; **b**, **d**, **f**) from ZT8 to ZT16 on day 10. Blue and red areas behind the traces represent blue and red periods, respectively. **g**, **h**, Fluence rate response curves for *CO* expression. LD-entrained plants were irradiated with different intensities of blue light (**g**) or red light (**h**) from ZT8 and harvested at ZT13 on day 10. Western blot images of FKF1-TAP and *CO* mRNA expression are shown. **i**, **j**, *CO* mRNA expression in *cry1 cry2* (**i**) and *phot1 phot2* (**j**) mutants in LD. For comparison with *cry1 cry2*, the *Ler* ecotype of wild type was used. Other features are the same as described in Fig. 2.

The light-induced difference spectra of the FKF1 LOV domain are consistent with those observed in formation of the long-lived photo-intermediate of the phototropin LOV domains¹⁶. This photo-intermediate involves the formation of a cysteinyl adduct between a protein cysteine group and the C4a carbon of the FMN chromophore¹⁷. Modelling the FKF1 LOV domain on the phytochrome3 LOV2 crystal structure (by using the Swiss model) places Cys 91 in close proximity to the FMN C4a carbon¹⁸. We found that replacing this conserved cysteine with an alanine abolished all photochemistry (Fig. 3a, c). Taken together, these results show that FKF1 possesses properties of a blue-light photoreceptor and may be able to perceive blue light. We also examined the ZTL LOV and LKP2 LOV domains and observed that they possess spectral properties similar to FKF1 LOV (Supplementary Fig. 5), indicating that FKF1, ZTL and LKP2 are likely to form a previously unknown family of photoreceptors in *Arabidopsis*.

As FKF1 might act as a blue-light photoreceptor, we examined the impact of light quality and quantity on FKF1 function. Plants were entrained in LD cycles, and on the final day of incubation were shifted to blue light or red light from ZT8 to ZT16, the critical time period of peak FKF1 amounts. In both light conditions, *CO* expression was affected in the *fkf1* mutant, indicating that FKF1 may be involved in both blue- and red-light regulation of *CO* (Fig. 4c, d, g, h). Under high intensities (greater than $25 \mu\text{mol m}^{-2} \text{s}^{-1}$) of monochromatic light, however, there was a much larger difference in *CO* induction between wild-type and *fkf1-2* plants in blue light than in red light, even though FKF1 was expressed in similar amounts in both conditions (Fig. 4a–d, g, h). Blue light was also effective in inducing *FT* expression in wild-type plants (Fig. 4e). Red light suppressed *FT* expression in both wild-type and *fkf1-2* plants (Fig. 4f). Thus, blue light seems to have a predominant role in activating FKF1 function, and this in turn induces expression of *CO* during the critical afternoon-to-dusk hours, when *CO* is important for *FT* induction.

Because blue light is a chief effector of FKF1 function, we examined whether FKF1 itself, or a different photoreceptor, mediates the blue-light perception required for *CO* induction. There are four known blue-light photoreceptors in *Arabidopsis*, two cryptochromes (*cry1* and *cry2*) and two phototropins (*phot1* and *phot2*)¹⁹. In addition, the red/far-red light photoreceptor phytochrome A (*phyA*) can also perceive blue light¹⁹. We reasoned that if the blue-light-dependent function of FKF1 relies on these photoreceptors, then the *CO* waveform in *cry*, *phot* or *phyA* mutants should be similar to that observed in *fkf1*. The daytime peak of *CO* in *cry1 cry2*, *phot1 phot2* and *phyA* mutants closely resembled the wild-type waveform (Fig. 4i, j, and Supplementary Fig. 6), indicating that FKF1 is not functioning downstream of these photoreceptors (although *FT* expression was altered in those mutants; see Supplementary Fig. 6). In conjunction with the spectral analysis, these results indicate that FKF1 may directly perceive and transduce the blue-light signal involved in the induction of *CO*.

It has been proposed that the precise timing of *CO* and *FT* expression is an essential factor underlying the photoperiodic control of flowering in *Arabidopsis*^{1,14}. Although the *cry2* and *phyA* photoreceptors have been shown to be key regulators of *FT* induction¹, little was previously known about the components regulating *CO*. Our results suggest that a crucial aspect of *CO* expression involved in day-length discrimination may be generated by the circadian control of FKF1 expression and by the light-dependent function of FKF1. In addition, we have provided biochemical and genetic evidence showing that FKF1 functions as a photoperiodic photoreceptor. Furthermore, our findings suggest that FKF1, ZTL and LKP2 constitute a family of blue-light photoreceptors in *Arabidopsis*. Together with previous studies, our work provides a more complete framework to explore the photoperiodic control of flowering in *Arabidopsis* (Supplementary Fig. 7). □

Methods

Plant materials

Except where indicated, the Colombia-0 (Col-0) ecotype was used as the wild type. *fkf1* (ref. 2), *cca1-1 lhy-R* (48)¹², *cry1 cry2* (ref. 20), *phot1-5 phot2-1* (ref. 21) and *phyA-201* (ref. 20) have been described. The *fkf1-2* T-DNA insertion line (593_H05) was obtained from the SAIL collection²⁴. The *cab2::luc* and *cer2::luc* reporters and 35S::CO (ref. 23) were introduced into *fkf1* alleles by crossing. To generate 35S::FKF1-TAP and 35S::LKP2-TAP, the coding regions of *FKF1* cDNA and *LKP2* cDNA were cloned into pRTL2-TAP (pRTL2 (ref. 24) containing a TAP²⁵ fragment) and then subcloned into pBI221 binary vector²⁶. To generate the *FKF1* promoter driven FKF1-TAP construct, we replaced the CaMV35S promoter in p35S::FKF1-TAP by 1.6 kb of the *FKF1* promoter plus the 5' untranslated region fragment. Unless otherwise noted, growth and light conditions were the same as described^{1,5}.

Analysis of gene expressions

In all experiments, plants were harvested on day 10 unless otherwise noted. RNA isolation and CO, FT and UBG quantification have been described^{1,27}. For *FKF1* expression analysis, total RNA (10 µg) was loaded on each lane. We used a 5' region (0.7 kb) of the *FKF1* cDNA as a probe. All traces of the gene expressions are the means ± s.e.m. of relative values of signals from three independent experiments.

Immunoblot analysis

Soluble proteins were extracted with buffer containing 50 mM HEPES-KOH (pH 7.5), 100 mM KCl, 10% glycerol, 5 mM EDTA, 2 mM dithiothreitol, 5% polyvinylpyrrolidone and Complete protease inhibitor cocktail tablets (Roche), and 120 µg of extract was loaded on each lane. We used peroxidase anti-peroxidase (Sigma) and anti-α-tubulin (Sigma) to detect the TAP-tagged proteins and tubulin, respectively. All immunoblot analyses were done twice on independent samples.

Analysis of flowering time

Plants were grown on MS agar plates with 3% sucrose for 2 weeks, and then transferred to soil in LD or SD. Flowering time was measured by counting both rosette and cauline leaves after bolting, when the inflorescence was 1-cm high.

Purification and analysis of LOV domain proteins

The LOV domains of FKF1 and the C91A mutant (amino acids 6–205), ZTL (1–190), LKP2 (5–194) and phot1 (449–599) were expressed in *Escherichia coli* as glutathione S-transferase (GST) fusion proteins and purified by GST affinity chromatography under safe (dim red) light. Roughly 150 µg of purified protein was used for thin layer chromatography analysis as described³. We measured absorbance spectra and difference spectra for the LOV domain fusions as described²⁸.

Received 20 June; accepted 22 September 2003; doi:10.1038/nature02090.

- Yanovsky, M. J. & Kay, S. A. Molecular basis of seasonal time measurement in *Arabidopsis*. *Nature* **419**, 308–312 (2002).
- Nelson, D. C., Lasswell, J., Rogg, L. E., Cohen, M. A. & Bartel, B. *FKF1*, a clock-controlled gene that regulates the transition to flowering in *Arabidopsis*. *Cell* **101**, 331–340 (2000).
- Christie, J. M., Salomon, M., Nozue, K., Wada, M. & Briggs, W. R. LOV (light, oxygen, or voltage) domains of the blue-light photoreceptor phototropin (nph1): binding sites for the chromophore flavin mononucleotide. *Proc. Natl Acad. Sci. USA* **96**, 8779–8783 (1999).
- Yanovsky, M. J. & Kay, S. A. Living by the calendar: how plants know when to flower. *Nature Rev. Mol. Cell Biol.* **4**, 265–275 (2003).
- Somers, D. E., Schultz, T. F., Milnamow, M. & Kay, S. A. *ZEITLUPE* encodes a novel clock-associated PAS protein from *Arabidopsis*. *Cell* **101**, 319–329 (2000).
- Schultz, T. F., Kiyosue, T., Yanovsky, M., Wada, M. & Kay, S. A. A role for LKP2 in the circadian clock of *Arabidopsis*. *Plant Cell* **13**, 2659–2670 (2001).
- Millar, A. J., Short, S. R., Chua, N. H. & Kay, S. A. A novel circadian phenotype based on firefly luciferase expression in transgenic plants. *Plant Cell* **4**, 1075–1087 (1992).
- Strayer, C. *et al.* Cloning of the *Arabidopsis* clock gene *TOC1*, an autoregulatory response regulator homolog. *Science* **289**, 768–771 (2000).
- Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17**, 1030–1032 (1999).
- Harmer, S. L. *et al.* Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science* **290**, 2110–2113 (2000).
- Alabadi, D. *et al.* Reciprocal regulation between *TOC1* and *LHY/CCA1* within the *Arabidopsis* circadian clock. *Science* **293**, 880–883 (2001).
- Alabadi, D., Yanovsky, M. J., Mas, P., Harmer, S. L. & Kay, S. A. Critical role for CCA1 and LHY in maintaining circadian rhythmicity in *Arabidopsis*. *Curr. Biol.* **12**, 757–761 (2002).
- Mizoguchi, T. *et al.* *LHY* and *CCA1* are partially redundant genes required to maintain circadian rhythms in *Arabidopsis*. *Dev. Cell* **2**, 629–641 (2002).
- Suarez-Lopez, P. *et al.* *CONSTANS* mediates between the circadian clock and the control of flowering in *Arabidopsis*. *Nature* **410**, 1116–1120 (2001).
- Christie, J. M. *et al.* *Arabidopsis* NPH1: a flavoprotein with the properties of a photoreceptor for phototropism. *Science* **282**, 1698–1701 (1998).
- Sakai, T. *et al.* *Arabidopsis* nph1 and npl1: blue light receptors that mediate both phototropism and chloroplast relocation. *Proc. Natl Acad. Sci. USA* **98**, 6969–6974 (2001).
- Salomon, M., Christie, J. M., Knieb, E., Lempert, U. & Briggs, W. R. Photochemical and mutational analysis of the FMN-binding domains of the plant blue light receptor, phototropin. *Biochemistry* **39**, 9401–9410 (2000).
- Crosson, S. & Moffat, K. Structure of a flavin-binding plant photoreceptor domain: insights into light-mediated signal transduction. *Proc. Natl Acad. Sci. USA* **98**, 2995–3000 (2001).
- Lin, C. Blue light receptors and signal transduction. *Plant Cell* **14** (suppl.), S207–S225 (2002).

- Yanovsky, M. J., Mazzella, M. A., Whitelam, G. C. & Casal, J. J. Resetting of the circadian clock by phytochromes and cryptochromes in *Arabidopsis*. *J. Biol. Rhythms* **16**, 523–530 (2001).
- Kinoshita, T. *et al.* phot1 and phot2 mediate blue light regulation of stomatal opening. *Nature* **414**, 656–660 (2001).
- Sessions, A. *et al.* A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell* **14**, 2985–2994 (2002).
- Onouchi, H., Igeno, M. I., Perilleux, C., Graves, K. & Coupland, G. Mutagenesis of plants overexpressing *CONSTANS* demonstrates novel interactions among *Arabidopsis* flowering-time genes. *Plant Cell* **12**, 885–900 (2000).
- Carrington, J. C., Freed, D. D. & Leinicke, A. J. Bipartite signal sequence mediates nuclear translocation of the plant potyviral N1a protein. *Plant Cell* **3**, 953–962 (1991).
- Puig, O. *et al.* The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218–229 (2001).
- Hajdukiewicz, P., Svab, Z. & Maliga, P. The small, versatile pPZP family of *Agrobacterium* binary vectors for plant transformation. *Plant Mol. Biol.* **25**, 989–994 (1994).
- Blazquez, M. A. & Weigel, D. Independent regulation of flowering by phytochrome B and gibberellins in *Arabidopsis*. *Plant Physiol.* **120**, 1025–1032 (1999).
- Swartz, T. E. *et al.* The photocycle of a flavin-binding domain of the blue light photoreceptor phototropin. *J. Biol. Chem.* **276**, 36493–36500 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Schultz, P. Más, F. Harmon and S. Hazen for critically reading the manuscript; Syngenta for the T-DNA insertion line; B. Bartel for *fkf1* and the pGEX FKF1 LOV construct; T. Kagawa and M. Wada for *phot1 phot2*; M. Yanovsky for *cry1 cry2*; G. Coupland for 35S::CO; and J. Harper for the TAP tag construct. This work was supported by grants from the NIH (to S.A.K. and H.G.T.); a grant from the NSF (to W.R.B.); and a grant from JSPS Postdoctoral Fellowships for Research Abroad (to T.I.). This is manuscript 15868-CB of The Scripps Research Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.A.K. (stevek@scripps.edu).

A genetic basis for *Pseudomonas aeruginosa* biofilm antibiotic resistance

Thien-Fah Mah¹, Betsey Pitts², Brett Pellock^{3*}, Graham C. Walker³, Phillip S. Stewart² & George A. O'Toole¹

¹Department of Microbiology and Immunology, Dartmouth Medical School, Hanover, New Hampshire 03755, USA

²Center for Biofilm Engineering, Montana State University, Bozeman, Montana 59717, USA

³Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02412, USA

* Present address: Massachusetts General Hospital Cancer Center, Building 149, 13th Street, Charlestown, Massachusetts 02129, USA

Biofilms are surface-attached microbial communities with characteristic architecture and phenotypic and biochemical properties distinct from their free-swimming, planktonic counterparts¹. One of the best-known of these biofilm-specific properties is the development of antibiotic resistance that can be up to 1,000-fold greater than planktonic cells². We report a genetic determinant of this high-level resistance in the Gram-negative opportunistic pathogen, *Pseudomonas aeruginosa*. We have identified a mutant of *P. aeruginosa* that, while still capable of forming biofilms with the characteristic *P. aeruginosa* architecture, does not develop high-level biofilm-specific resistance to three different classes of antibiotics. The locus identified in our screen, *ndvB*, is required for the synthesis of periplasmic glucans. Our discovery that these periplasmic glucans interact physically with tobramycin suggests that these glucose polymers may prevent antibiotics from reaching their sites of action by seques-

tering these antimicrobial agents in the periplasm. Our results indicate that biofilms themselves are not simply a diffusion barrier to these antibiotics, but rather that bacteria within these microbial communities employ distinct mechanisms to resist the action of antimicrobial agents.

Surface-attached bacterial communities, or biofilms, can become 10–1,000 times more resistant to the effects of antimicrobial agents than their planktonic counterparts². This characteristic of biofilms makes them extremely difficult to control in medical and industrial settings³. We hypothesized that biofilm resistance to antimicrobial agents is part of a regulated developmental process and thus would require an identifiable set of genetic determinants.

Using a 96-well microtitre plate-based assay for biofilm-specific antibiotic resistance, a library of ~4,000 random *P. aeruginosa* PA14 transposon insertion mutants was screened for the inability to develop the characteristic increase in resistance of wild-type biofilm-grown cells. A mutant was identified, designated 45E7, that had decreased resistance to tobramycin (Tb) when grown in a biofilm, but grew as well as the wild type in planktonic culture, formed a biofilm indistinguishable from a biofilm formed by the wild-type strain in the microtitre plate assay, and displayed a planktonic minimal bactericidal concentration (MBC) for Tb that was indistinguishable from the parent strain.

To assess the generality of biofilm antibiotic sensitivity of the mutant, the MBC of biofilm-grown and planktonically grown wild-type and 45E7 strains was determined for gentamicin (Gm), ciprofloxacin (Cip), nalidixic acid (Nal), chloramphenicol (Cm) and ofloxacin (Of). For all of these antibiotics tested, there was no difference in the planktonic MBC between the wild type and the 45E7 mutant. The MBC of biofilm-grown 45E7 was eightfold lower for Gm and Cip, twofold lower for Cm and Of, and unchanged for Nal when compared to the wild type (Table 1 and see Supplementary Information). Therefore, the 45E7 locus can account for some, but not all, of the biofilm-mediated resistance observed for most antibiotics.

We used three other assays of biofilm antibiotic resistance to confirm the antibiotic sensitivity phenotype observed in the microtitre plate. The Tb-sensitivity phenotype in a flow cell was demonstrated by treating a 24-h-old, pre-formed biofilm with Tb, then staining the Tb-treated biofilm with the BacLight viability stain (Fig. 1). When viewed by phase-contrast microscopy (Fig. 1a, left panels), no difference in biofilm architecture was discernible between the wild type and the mutant (see also Table 2). The centre and right panels show that there were more live (green, syto-9-stained) than dead (red, stained with propidium iodide) cells in the wild-type biofilm after treatment with Tb. Conversely, there were more dead cells than live ones in the 45E7 biofilm. Quantification of propidium iodide-dependent fluorescence demonstrated 100-fold greater fluorescence intensity for the 45E7 mutant versus the wild type. When 72-h-old biofilms were treated with Tb, the 45E7 mutant biofilm was only threefold more sensitive than the wild-type biofilm (see Supplementary Information), suggesting that the 45E7 locus plays a more important role in younger biofilms. As a control, biofilms were stained with BacLight before exposure to Tb (Fig. 1b) and most cells in both the wild-type and 45E7 biofilms stained green, indicating that the 45E7 mutant does not have a

general viability defect.

We used a second, quantitative assay to document the 45E7 mutant phenotype. The colony biofilm assay exploits the observation that bacterial colonies develop some of the properties of biofilms, including increased antibiotic resistance^{4,5} (Fig. 2). 45E7 biofilm cells remained as resistant to the effects of Tb as the wild type at the first two time points assayed. However, at 48 h, while the wild-type biofilm cell viability was only reduced by two orders of magnitude, the viable cells detected for the 45E7 mutant decreased by at least five orders of magnitude. There was no difference in the viability of the strains in the absence of antibiotics over this time course. Furthermore, we found that there was no difference in the ability of Tb to diffuse through the colony biofilms formed by either of these strains (data not shown).

The rotating disk assay⁶ was employed as another quantitative method to assess the phenotype of the 45E7 mutant. After forming on coupons, wild-type and mutant biofilms were exposed to varying concentrations of Tb. A lower concentration of Tb was necessary to kill the cells of the mutant biofilm compared to the wild-type biofilm (see Supplementary Information). These results, taken together with the data presented above, indicate conclusively that the 45E7 mutant, when growing in a biofilm, is less resistant than the wild type to the antimicrobial effects of Tb.

Next, we used flow cells to analyse the architecture of the wild-type and mutant strains. Both strains formed heterogeneous macrocolonies and fluid-filled channels that are characteristic of *P. aeruginosa* biofilms⁷ (Fig. 1). To quantitate the architecture of biofilms formed by these strains, we used the COMSTAT software program⁷. COMSTAT analysis showed that there is no discernible difference between these strains for a number of architectural parameters at 24 h (Table 2).

The transposon in the 45E7 strain was mapped to the middle of a

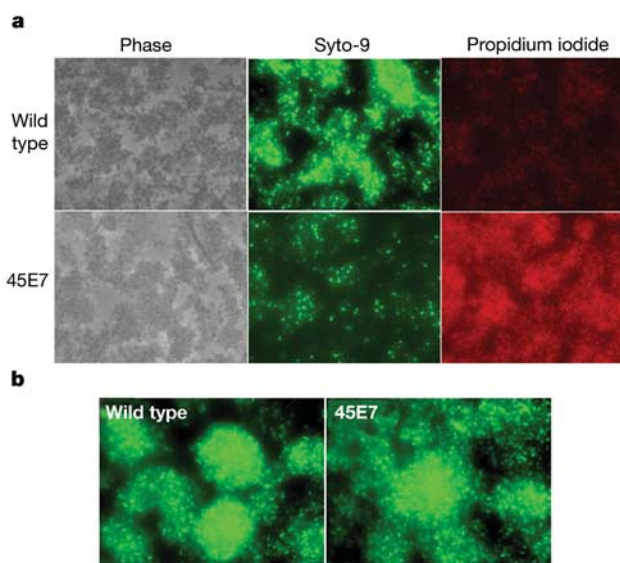


Figure 1 Flow cell assay for antibiotic sensitivity. **a**, The left-most panels (phase-contrast micrographs) show the architecture of 24-h-old biofilms. A flow cell allows a continuous supply of fresh medium to be delivered to a biofilm that is formed on the walls of a small, enclosed chamber. Biofilms were allowed to pre-form for 24 h in the absence of antibiotic treatment. These 24-h-old biofilms were then treated with 20 $\mu\text{g ml}^{-1}$ of Tb for 24 h, followed by staining with the BacLight viability stain for 1 h, then a 1 h rinse to remove unincorporated stain. The syto-9 panels indicate viable cells (green, intact membranes) and the propidium iodide panels indicate dead cells (red, damaged membranes). Similar results were observed in three separate experiments. **b**, Syto-9 staining of the wild-type and 45E7 biofilm before Tb treatment revealed no difference in viability.

Table 1 Antibiotic sensitivity of biofilm-grown and planktonically grown strains

Strain	Tb		Gm		Cip	
	MBC-P	MBC-B	MBC-P	MBC-B	MBC-P	MBC-B
Wild type	8	400	40	500	4	50
45E7	8	25	40	60	4	6
Fold change	1 $\times$	16 $\times$	1 $\times$	8 $\times$	1 $\times$	8 $\times$

MBC-P, MBC of planktonically grown cells; MBC-B, MBC of biofilm-grown cells. All antibiotic concentrations are in $\mu\text{g ml}^{-1}$. Values are based on results from at least three separate experiments.

monocistronic open reading frame (ORF), PA1163⁸, that is 58% identical across its entire sequence to the *Bradyrhizobium japonicum* gene *ndvB*. The *ndvB* gene codes for a glucosyltransferase that is required for the synthesis of cyclic- β -(1, 3)-glucans⁹. Located in the periplasm and also secreted into the extracellular media, cyclic glucans are circular polymers of glucose that have been shown to play a role in adaptation to low osmotic media, flagella-mediated motility and plant-microbe interactions¹⁰. Furthermore, a planktonically grown *B. japonicum* strain carrying an *ndvB* mutation is more sensitive to chloramphenicol than the parent strain¹¹.

To confirm that mutation in the *ndvB* gene was causative for the observed phenotypes, we disrupted the gene in *P. aeruginosa* strains PA14, PAO1 and PAK, and tested the antibiotic sensitivity of these mutants in the 96-well microtitre plate assay. In all cases, the mutant was more sensitive than the wild-type strain when growing in a biofilm, but no differences were observed between the strains grown planktonically (see Supplementary Information). However, in the two additional strains, the influence of *ndvB* was not as great as for PA14. Most notably, only a twofold difference was observed for PAO1. Although the data support an involvement for *ndvB* in biofilm antibiotic resistance, it may be that some strains of *Pseudomonas* have compensatory mechanisms for *ndvB* mutations.

To determine whether a difference in periplasmic glucan production existed between the wild type and the *ndvB* mutant, we examined ethanol-extracted periplasmic material⁹ purified from both strains by gel filtration over a Sephadex G-75 column (detected using the anthrone-sulphuric acid method¹²). The most striking difference between wild-type and mutant chromatograms was observed between fractions 60 and 68 (Fig. 3a). Whereas a significant amount of anthrone-positive material was present in these fractions in the wild type, the equivalent fractions from the *ndvB* mutant contained relatively little anthrone-positive material. Because periplasmic cyclic glucan material from *Sinorhizobium meliloti* eluted from the same column in the same area (the main peak at fraction 62; data not shown), we reasoned that the missing material from the mutant 45E7 might be periplasmic glucans. Monosaccharide composition analyses using Dionex high-performance anion exchange chromatography coupled with pulsed amperometric detection (HPAEC-PAD)¹³ indicated that the material present in fractions 60–68 from the wild type, and missing from the mutant, was composed exclusively of glucose (see Supplementary Information). Taken together, our genetic and carbohydrate

analyses strongly suggest that the *ndvB* mutant is deficient in production of periplasmic glucans.

On the basis of the current understanding of the roles of periplasmic glucans in *B. japonicum*, *S. meliloti* and *Agrobacterium tumefaciens*, we envisioned several possible models to account for the role of these molecules in development of resistance in biofilm populations. Unlike the effects of *ndvB* mutations in Rhizobacteria¹⁰, the *P. aeruginosa ndvB* mutant had no apparent defects in adaptation to low osmotic stress conditions or flagella-mediated motility. From the observation that related cyclic sugar polymers called cyclodextrins could sequester “guest molecules” in a non-covalent fashion¹⁴, we considered the possibility that one role of periplasmic glucans may be to sequester antimicrobial agents in biofilm-grown cells¹⁰. If periplasmic glucans were indeed sequestering Tb, we predicted that these molecules should interact *in vitro*. To test this, we examined interactions of crude and purified glucans with Tb by hydrophobic interaction (HI) chromatography. Tb alone was loaded onto a HI interaction resin (C18) column, and was detected in the column flow-through and first water wash of the

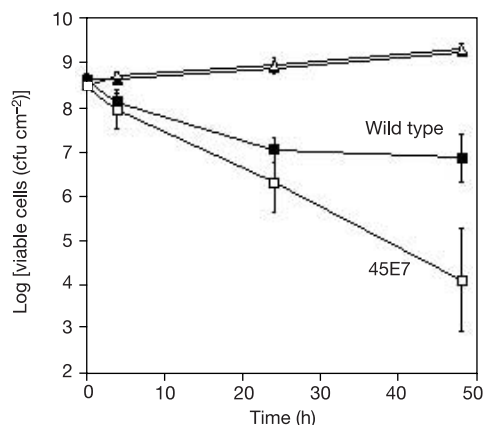


Figure 2 Colony biofilm assay for antibiotic sensitivity. 48-h-old colonies of the wild-type (filled symbols) and the 45E7 mutant (open symbols) were transferred to solid medium without (triangles) or with 10 $\mu\text{g ml}^{-1}$ Tb (squares). The viability of the wild type decreased 100-fold with Tb treatment relative to the untreated control, while the 45E7 mutant decreased 10⁵-fold. In the absence of antibiotic, there was no difference in viability between the two strains.

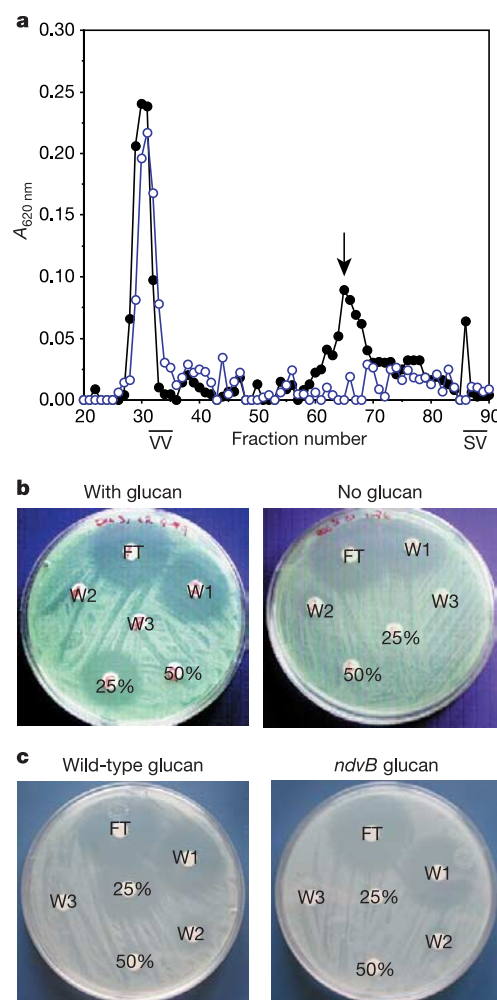


Figure 3 Periplasmic glucans interact with Tb. **a**, Gel filtration chromatography profiles of the periplasmic extracts of wild type (filled circles) and *ndvB* mutant (open circles). The arrow indicates the periplasmic glucan peak. vv, void volume; sv, salt volume. **b**, Disk diffusion assays with and without pre-loaded crude wild-type periplasmic glucans. Eluates from C18 columns were spotted on a lawn of Tb-sensitive *E. coli*. FT, column flow-through; W1–W3, water washes; 25%, 25% acetonitrile fractions; 50%, 50% acetonitrile fractions. The 75% and 100% acetonitrile fractions did not contain any Tb (data not shown). **c**, Disk diffusion assay with pre-loaded wild-type-derived (left) and *ndvB*-derived (right) fractions 60–68. Disks are labelled as above.

Table 2 Quantifying biofilm architecture

Strain	Total biomass ($\mu\text{m}^3 \mu\text{m}^{-2}$)	Average thickness (μm)	Roughness coefficient	Surface area covered (μm^2)	Surface/volume ($\mu\text{m}^2 \mu\text{m}^{-3}$)	Max thickness (μm)
Wild type	4.98 (3.18)	5.08 (3.24)	1.05 (0.31)	1.2×10^7 (4.9×10^6)	0.30 (0.10)	17.67 (2.47)
45E7	7.05 (3.64)	7.33 (3.80)	0.79 (0.38)	1.8×10^7 (7.4×10^6)	0.33 (0.09)	18.46 (2.01)

Note. Standard deviation (s.d.) for each data set is shown in parentheses. These results are an average of 12 data sets and all differences are within error. Roughness coefficient is a measure of biofilm heterogeneity.

column, but not in any subsequent fractions (Fig. 3b, right). In contrast, when a crude periplasmic carbohydrate extract from the wild-type strain was pre-loaded onto the column, some of the Tb was retained on the column and eluted with 25% acetonitrile (Fig. 3b, left). Conversely, when crude extract from the *ndvB* mutant strain was pre-loaded onto the column, very little Tb was retained on the column (see Supplementary Information).

To determine which components of the crude extract were responsible for the Tb retention on the C18 column, we performed the same experiment as above except the C-18 column was pre-loaded with partially pure material from fractions 60–68 from G-75 columns (Fig. 3c). Tb was retained on the column that was pre-loaded with material from the wild-type extract and eluted from the column with 25% acetonitrile. In contrast, no Tb activity was present in the 25% acetonitrile fraction from the column that was pre-loaded with the corresponding fractions isolated from the *P. aeruginosa ndvB* mutant. These data support the hypothesis that periplasmic glucans may sequester antibiotics, thereby slowing or stopping the diffusion of these molecules to their site of action.

On the basis of the biofilm-specific role of *ndvB* in resistance, we predicted that this gene is expressed only in biofilm-grown cells. RNA was prepared from biofilm-grown and planktonically grown cells from a continuous culture system¹⁵. Semi-quantitative RT–PCR using primers to *ndvB* revealed that there was preferential expression of *ndvB* in the biofilm mode of growth (Fig. 4).

We propose a mechanism for biofilm-specific resistance wherein periplasmic glucans bind to and sequester antibiotics. This extra-cytoplasmic mode of antibiotic sequestration would be expected to interfere with the passage of antibiotics through the periplasmic space to their site of action in the cytoplasm. High levels of the

antibiotic needed to eliminate biofilm bacteria may be required to overwhelm the ability of the glucans to bind antimicrobial agents effectively. Alternatively, periplasmic glucans may not entirely account mechanistically for all of the antibiotic resistance of biofilm cells but may contribute to it by slowing the diffusion of antibiotics into the cell, thereby allowing the bacteria additional time to adapt to the antibiotic insult, an idea consistent with the model proposed previously¹⁵.

We also showed that the *P. aeruginosa ndvB* mutant had increased biofilm-sensitivity to five antibiotics from three different structural families. This observation is consistent with previous reports that cyclic glucans can bind a range of chemically distinct molecules¹⁰ and the reported ability of biofilms to develop broad resistance to antimicrobial agents¹⁶. It may now be possible to identify new strategies to block the development of biofilm resistance. For example, a co-therapeutic approach where traditional antibiotics are combined with a drug that interferes with biofilm-specific resistance may render biofilms more susceptible to treatment. □

Methods

Bacterial strains, plasmids, media and chemicals

P. aeruginosa PA14 was grown on rich medium (Luria Bertani, LB) or minimal medium at 37 °C. The minimal medium used was minimal M63 salts¹⁷ supplemented with arginine (0.4%) and MgSO_4 (1 mM). The pSMC21 (green fluorescent protein, GFP) plasmid used to mark cells for flow cell studies is a kanamycin-resistant derivative of the reported pSMC2 plasmid¹⁸. We generated pTFM1 by ligating a 2-kilobase *ndvB* fragment into the *Bam*H1 and *Eco*R1 sites of pEX18Gm. The *ndvB* fragment contained 650 nucleotides of 5' *ndvB* sequence, generated using primers p242: 5'-GGCGGATCCCAACCCTCTGAAACCGAATCGCGGATTG-3' and p243: 5'-GGCGGCGCATATGGGCTTCGATGACG AAGTAGCTGTAGCCTTC-3' in a standard polymerase chain reaction (PCR) reaction, and 650 nucleotides of 3' *ndvB* sequence, generated using the primers p244: 5'-GGCGG CGCATATGCTGTTCGCTT CAAGGTCGGCAAGATCCTC-3' and p245: 5'-GGCGG AATTCGAGTTCAGGTTG GAAACCTGGAAGCTGGTC-3', in a standard PCR reaction. The BacLight Live/Dead viability stain was purchased from Molecular Probes.

Genetic and molecular techniques

Transposon mutants were generated with Tn5-B30¹⁹ as described²⁰. To isolate strains defective in biofilm antibiotic resistance, individual transposon mutants were inoculated from cultures grown overnight in LB into sterile 96-well microtitre plates containing minimal arginine medium. Biofilms were allowed to form for 24 h, after which time the medium was replaced with fresh minimal arginine medium supplemented with $50 \mu\text{g ml}^{-1}$ Tb (Research Products International Corporation). After 24 h of exposure to Tb, biofilms were allowed to recover in fresh media lacking Tb for an additional 24 h; their viability was then assessed by plating a small volume of each culture on LB plates. The minimal bactericidal concentration for biofilm-grown cells (MBC-B) was determined by exposing 24-h-old biofilms to various concentrations of antibiotic for 24 h, followed by a 24 h recovery period and plating for viability. The MBC of planktonic bacteria (MBC-P) was determined by adding antibiotic to bacteria at the time of inoculation, incubating for 24 h, followed by plating for viability. In these experiments, the number of bacteria used in planktonic cultures and microtitre plate-grown biofilms was roughly equivalent ($\sim 10^7$ c.f.u. per well). DNA sequence flanking the transposon mutant 45E7 was determined as described^{20,21}. The original screen of $\sim 4,000$ transposon mutants yielded two mutants with the desired phenotype, one of which is presented in this report (see text). The *ndvB* deletion strains of PA14, PA01 and PAK were generated using pTFM1 as reported²².

Antibiotic resistance assays

Flow cells were prepared, sterilized and analysed as described²³. Image acquisition and quantification of fluorescence intensity were performed with OpenLab software using the Morphology Module (Improvision). Quantification of biofilm architecture was performed with the COMSTAT software package as described²⁴. Colony biofilms^{4,5} and rotating disk experiments⁶ were performed as reported.

Periplasmic glucan purification

Periplasmic glucans were purified as described⁹. Briefly, 750 ml cultures of wild-type and

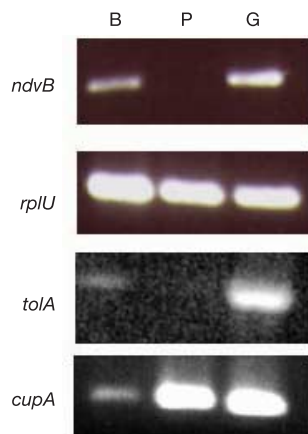


Figure 4 *ndvB* is preferentially expressed in biofilms. RT–PCR analysis of *ndvB* expression. PCR products derived from RNA isolated from planktonic (P) and biofilm (B) cells were analysed by agarose gel electrophoresis. A control with added genomic DNA (G) shows the expected size of the band with the primer set used. The constitutively expressed *rplU* gene was used as a control in the experiment. The similar bands from P and B samples indicate that an equal amount of RNA was present in each sample. *tolA* and *cupA* transcripts were used as biofilm and planktonic controls¹⁵, respectively. When RNA, but not cDNA, was used in the reaction, no bands were detected (data not shown).

ndvB strains were grown for 24 h at 37 °C in arginine minimal medium. The cells were pelleted and the glucans were extracted with 70% EtOH at 70 °C for 30 min. Cell debris was pelleted and glucans were precipitated with ten volumes of 100% EtOH and 100 mM NaCl. Glucans were pelleted and resuspended in water. Gel filtration chromatography of the crude periplasmic extract was performed with Sephadex G-75 resin as described²⁵.

Tobramycin–glucan interaction assays

C18 Sep-Pak columns (Waters) were pre-washed with two column volumes of 50% methanol, followed by two column volumes of water. Washed columns were pre-loaded with 275 µg of resuspended glucans, or an equal volume (250 µl or 2.5% of the extract) of the crude wild-type or *ndvB* periplasmic extracts, washed with two column volumes of water then loaded with 1 mg of Tb. Following water washes (three column volumes), the column was eluted step-wise with one column volume each of 25%, 50%, 75% and 100% acetonitrile. Fractions were evaporated to dryness and resuspended in water. 12.5% of each sample was spotted on a sterile Whatman no. 1 paper disk on a lawn of *Escherichia coli* DH5α spread onto a LB plate. The lawn of *E. coli* was prepared by diluting an overnight culture 1:100, then spreading the dilution on a LB plate. Plates were incubated at 37 °C overnight before photographing. Tb-containing fractions were detected by the zone of inhibition surrounding the paper disk.

RT-PCR

Biofilm and planktonic RNA used in reverse transcriptase (RT)–PCR assays was purified with the RNeasy kit (Qiagen) from bacteria grown in continuous flow reactors as described¹⁵. Complementary DNA was synthesized using the first strand cDNA synthesis kit (Pharmacia) and the resulting RNA/cDNA hybrid was used as a template in PCR reactions using the primer sets designed from the *P. aeruginosa* PAO1 genome sequence⁸: *ndvB*: 5′-CGAACAGATGCGCACCGACC-3′ and 5′-CGCAGGTAGATGGCCTGGTC-3′; *rplU*: 5′-CGCAGTGATTGTTACCGGTG-3′ and 5′-AGGCCTGAATGCCGCTGATC-3′; *tolA*: 5′-CACGTTCTGATCTTCGCCATGCTG-3′ and 5′-GTCGGTCGTATCCGAAAGGAACTC-3′; *cupA1*: 5′-CATGCGCAGTGGTATTGGCCTTTG-3′ and 5′-GAACAGGGTGGTGAATGCTCGTC-3′. The PCR was performed by using Taq polymerase (Qiagen) for 20 cycles with an annealing temperature of 64 °C (*ndvB*) or 56 °C (*rplU*, *tolA* and *cupA1*) and an extension temperature of 72 °C. The resulting products were analysed on a 1.2% agarose gel.

Received 19 August; accepted 9 October 2003; doi:10.1038/nature02122.

- Davey, M. E. & O'Toole, G. A. Microbial biofilms: from ecology to molecular genetics. *Microbiol. Mol. Biol. Rev.* **64**, 847–867 (2000).
- Hoyle, B. D. & Costerton, W. J. Bacterial resistance to antibiotics: the role of biofilms. *Prog. Drug Res.* **37**, 91–105 (1991).
- Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: a common cause of persistent infections. *Science* **284**, 318–322 (1999).
- Anderl, J. N., Franklin, M. J. & Stewart, P. S. Role of antibiotic penetration limitation in *Klebsiella pneumoniae* biofilm resistance to ampicillin and ciprofloxacin. *Antimicrob. Agents Chemother.* **44**, 1818–1824 (2000).
- Walters, M. C., Roe, F., Bugnicourt, A., Franklin, M. J. & Stewart, P. S. Contributions of antibiotic penetration, oxygen limitation, and low metabolic activity to tolerance of *Pseudomonas aeruginosa* biofilms to ciprofloxacin and tobramycin. *Antimicrob. Agents Chemother.* **47**, 317–323 (2003).
- Hentzer, M. et al. Alginate overproduction affects *Pseudomonas aeruginosa* biofilm structure and function. *J. Bacteriol.* **183**, 5395–5401 (2001).
- Heydorn, A. et al. Quantification of biofilm structures by the novel computer program COMSTAT. *Microbiol.* **146**, 2395–2407 (2000).
- Stover, C. K. et al. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**, 959–964 (2000).
- Bhagwat, A. A., Gross, K. C., Tully, R. E. & Keister, D. L. β-glucan synthesis in *Bradyrhizobium japonicum*: characterization of a new locus (*ndvC*) influencing beta-(1→6) linkages. *J. Bacteriol.* **178**, 4635–4642 (1996).
- Breedveld, M. W. & Miller, K. J. Cyclic beta-glucans of members of the family Rhizobiaceae. *Microbiol. Rev.* **58**, 145–161 (1994).
- Chen, R., Bhagwat, A. A., Yaklich, R. & Keister, D. L. Characterization of *ndvD*, the third gene involved in the synthesis of cyclic β-(1→3), (1→6)-D-glucans in *Bradyrhizobium japonicum*. *Can. J. Microbiol.* **48**, 1008–1016 (2002).
- Loewus, F. A. Improvement in the anthrone assay for determination of carbohydrates. *Anal. Chem.* **24**, 219 (1952).
- Gonzalez, J. E., Semino, C. E., Wang, L. X., Castellano-Torres, L. E. & Walker, G. C. Biosynthetic control of molecular weight in the polymerization of the octasaccharide subunits of succinoglycan, a symbiotically important exopolysaccharide of *Rhizobium meliloti*. *Proc. Natl Acad. Sci. USA* **95**, 13477–13482 (1998).
- Szejtli, J. The cyclodextrins and their applications in biotechnology. *Carbohydr. Polym.* **12**, 375–392 (1990).
- Whiteley, M. et al. Gene expression in *Pseudomonas aeruginosa* biofilms. *Nature* **413**, 860–864 (2001).
- Mah, T.-F. & O'Toole, G. A. Mechanisms of biofilm resistance to antimicrobial agents. *Trends Microbiol.* **9**, 34–39 (2001).
- Pardee, A. B., Jacob, F. & Monod, J. The genetic control and cytoplasmic expression of “inducibility” in the synthesis of β-galactosidase in *E. coli*. *J. Mol. Biol.* **1**, 165–178 (1959).
- Bloemberg, G. V., O'Toole, G. A., Lugtenberg, B. J. J. & Kolter, R. Green fluorescent protein as a marker for *Pseudomonas* spp. *Appl. Environ. Microbiol.* **63**, 4543–4551 (1997).
- Simon, R., Quandt, J. & Klipp, W. New derivatives of transposon Tn5 suitable for mobilization of replicons, generation of operon fusions and induction of genes in Gram-negative bacteria. *Gene* **80**, 160–169 (1989).
- O'Toole, G. A. & Kolter, R. The initiation of biofilm formation in *Pseudomonas fluorescens* WCS365 proceeds via multiple, convergent signaling pathways: a genetic analysis. *Mol. Microbiol.* **28**, 449–461 (1998).

- Caetano-Annoles, G. Amplifying DNA with arbitrary oligonucleotide primers. *PCR Methods Appl.* **3**, 85–92 (1993).
- Donnenberg, M. S. & Kaper, J. B. Construction of an *eae* deletion mutant of enteropathogenic *Escherichia coli* by using a positive-selection suicide vector. *Infect. Immun.* **59**, 4310–4317 (1991).
- Christensen, B. B. et al. Molecular tools for study of biofilm physiology. *Methods Enzymol.* **310**, 20–42 (1999).
- Heydorn, A. et al. Experimental reproducibility in flow-chamber biofilms. *Microbiol.* **146**, 2409–2415 (2000).
- Wang, L. X., Wang, Y., Pelloch, B. & Walker, G. C. Structural characterization of the symbiotically important low-molecular-weight succinoglycan of *Sinorhizobium meliloti*. *J. Bacteriol.* **181**, 6788–6796 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L.-X. Wang for discussions. This work was supported by grants from the NSF to P.S.S., from the NIH to G.C.W., from the Canadian Cystic Fibrosis Foundation to T.-F.M. and from the NIH, Microbia, Inc. and The Pew Charitable Trusts to G.A.O.T., who is a Pew Scholar in the Biomedical Sciences.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to G.A.O.T. (georgeo@Dartmouth.edu).

Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi

Titia Sijen & Ronald H. A. Plasterk

Hubrecht Laboratory, Uppsalaalaan 8, 3584 CT, Utrecht, The Netherlands

Transposable elements are stretches of DNA that can move and multiply within the genome of an organism. The *Caenorhabditis elegans* genome contains multiple *Tc1* transposons that jump in somatic cells, but are silenced in the germ line^{1–3}. Many mutants that have lost this silencing have also lost the ability to execute RNA interference (RNAi)^{2,3}, a process whereby genes are suppressed by exposure to homologous double-stranded RNA (dsRNA). Here we show how RNAi causes transposon silencing in the nematode germ line. We find evidence for transposon-derived dsRNAs, in particular to the terminal inverted repeats, and show that these RNAs may derive from read-through transcription of entire transposable elements. Small interfering RNAs of *Tc1* were detected. When a germline-expressed reporter gene is fused to a stretch of *Tc1* sequence, this transgene is silenced in a manner dependent on functional mutator genes (*mut-7*, *mut-16* and *pk732*). These results indicate that RNAi surveillance is triggered by fortuitous read-through transcription of dispersed *Tc1* copies, which can form dsRNA as a result of ‘snap-back’ of the terminal inverted repeats. RNAi mediated by this dsRNA silences transposase gene expression.

Genome-sequencing projects have shown that genomes contain large numbers of transposable elements; for example, 45% of the human genome consists of remnants of transposon sequences, whereas 2% encodes protein. Transposon sequences comprise 12% of the *C. elegans* genome; of the several DNA transposons, *Tc1* is the most abundant element, with 31 intact copies⁴. Transposition (through a cut-and-paste mechanism) is mediated by a *Tc1*-encoded transposase that specifically recognizes the *Tc1* terminal inverted repeats (TIRs)⁵. Transposition only occurs in somatic cells and not in the germ line¹. The identification of mutator mutants, which do show germline transposition, indicated that an active transposon-silencing process exists in the germ line. Several

mutator mutants were found to be defective in RNAi^{2,3}, a gene-silencing mechanism in which dsRNA triggers sequence-specific RNA degradation. This raised the possibility that transposon silencing may be a natural function of RNAi⁶.

As RNAi is triggered by dsRNA, we analysed wild-type nematodes for the presence of transposon-derived dsRNA⁷. In RNase protection assays, we detect dsRNA predominantly using probes containing the Tc1 TIR (Fig. 1a). Moreover, the length of the protected fragment corresponds precisely to the number of TIR nucleotides (nt) present in the probe (Fig. 1a), showing that transposon dsRNA has its origin mainly in the TIR sequences. TIR-derived dsRNA is also detected for the transposable elements Tc3 and Tc5, both of which have longer TIRs and fewer genomic copies (data not shown). The finding that transposon dsRNA is limited to the TIR implies that transposon transcripts form dsRNA intramolecularly by snap-

ping back into a panhandle structure. TIR dsRNA is also observed in isolated oocytes (in a *fer-1* mutant background⁸; data not shown), indicating that in germline tissue (where transposon silencing occurs) transposon dsRNA is indeed present. Tc1 TIR dsRNA is also detected in several mutants (*mut-6*, *pk739*, *pk724* and *pk732*) defective only in transposon silencing and not in RNAi², indicating that transposon mobilization in these strains is not due to the absence of transposon dsRNA.

In *C. elegans*, as in many other organisms, dsRNA is prone to RNA editing by ADARs, enzymes that deaminate adenosine to create inosine⁹. To determine whether transposon dsRNA is present *in vivo* (rather than being an artefact of RNA isolation), we assayed the occurrence of RNA editing in wild-type worms by 5'-rapid amplification of cloned ends (RACE) analyses in which individual Tc1 reverse transcriptase products were cloned and sequenced. As inosine is read as guanosine, A to G (or T to C) transitions are indicative of RNA editing. RNA editing is observed primarily in the Tc1 TIR sequences and rarely in the flanking *C. elegans* sequences present in these 5'-RACE clones (Table 1). Analysis of non-TIR (internal) sequences of Tc1 reveals that significant editing occurs only for transcripts derived from locus *T10B5* (Table 1), which contains two Tc1 elements oriented in an inverted repeat—tail-to-tail orientation, spaced by 222 base pairs (bp). Thus, editing occurs mainly in RNA sequences that are able to undergo intramolecular base pairing. Furthermore, these data support the existence of transposon-derived dsRNA *in vivo*.

The presence of dsRNA over the full extent of the transposon TIR implies that transcription of the transposon sequence started in the flanking sequences and covered the complete transposable element. In agreement with this, primer-extension studies do not reveal an internal promoter in Tc1 (Fig. 1b), and 5'-RACE analyses in wild-type worms show that many of the 31 Tc1 loci are transcribed by read-through transcription from *C. elegans* genes in either sense or antisense orientation (Table 2). RT-PCR (polymerase chain reaction with reverse transcription) analysis using primers specific for the flanking sequences of Tc1 locus *ZK250* show the presence of a sense transcript containing the complete Tc1 sequence (Fig. 1c), which may 'snap back' and form dsRNA for the full extent of the TIR.

Next, we examined whether these read-through transcripts undergo normal RNA processing. We find that introns residing in the flanking sequences are spliced out (5'-RACE analyses); however, the Tc1 intron was only spliced out in a limited number of transcripts (data not shown). In RNase protection assays using an antisense probe covering the Tc1 intron sequence, we detect few (around 2%) spliced transcripts, whereas for an actin gene only spliced transcripts are found. Also, 5' *trans*-splicing is detected for transcripts from three Tc1 loci as apparent from a 5'-SL1 leader sequence (RT-PCR analyses). No polyadenylated Tc1 transcripts are found (RT-PCR analyses; data not shown). Apparently, few Tc1 messenger RNAs are correctly processed, although these may be

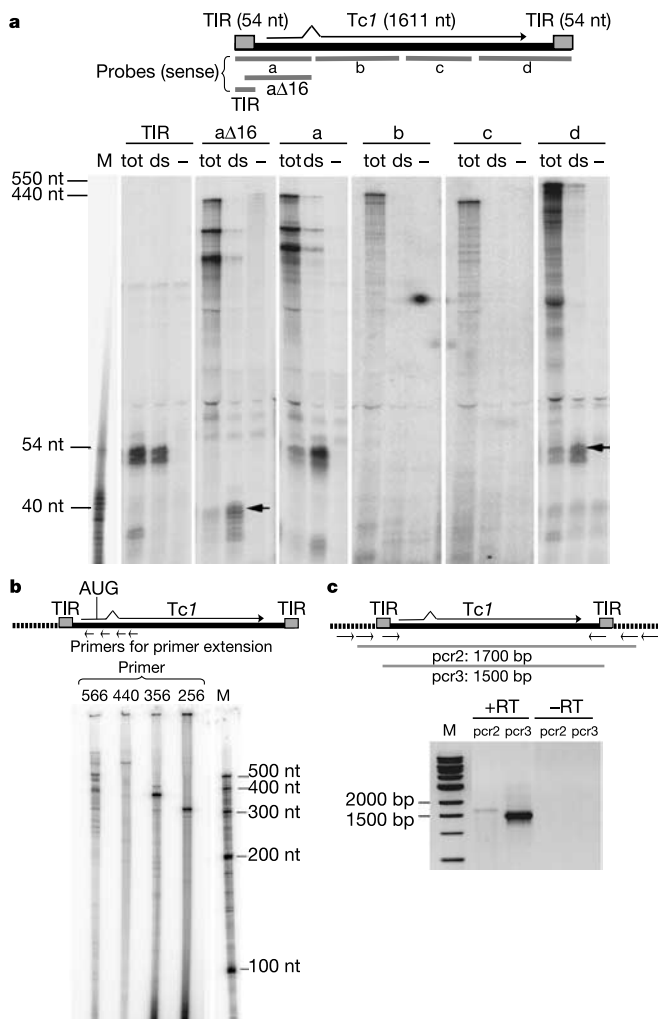


Figure 1 Analyses of Tc1 RNAs. **a**, Detection of dsRNA by RNase protection assays on RNA isolated from wild-type worms. For each probe, the RNA was either not pre-treated (tot), pre-treated with RNaseA in high salt (ds) (degradation of ssRNA only) or pre-treated with RNaseA in low salt (–) (degradation of both dsRNA and ssRNA). The position of the probes in Tc1 is indicated. M, RNA size marker. Protected fragments corresponding to dsRNA are indicated by an arrow. Similar protected fragments were detected when antisense probes were used (data not shown). **b**, Primer-extension analyses do not reveal an obvious transcription start site in Tc1 as the main extension products are larger than the distance from the primer position to nucleotide 1 in Tc1. The nucleotide position of each primer is indicated above each lane. AUG is at position 273. M, RNA size marker. **c**, RT-PCR analysis (twice nested) to detect sense transcripts covering the full-length Tc1 element at locus *ZK250*.

Table 1 RNA editing in Tc1 sequences

Analysed sequence	No. of clones edited	No. of nucleotides edited	No. of Tc1 loci edited
Tc1 TIR (30 nt)	15* (157)	32 (450)	4
Flanks (edited clones)	1* (15)	2 (783)	
Tc1 internal	7† (76)	58 (8,077)	2
Tc1 internal (<i>T10B5</i>)	5† (5)	55 (847)	

Number of edited nucleotides in various regions of Tc1-containing transcripts. The total number of clones or nucleotides checked is indicated in parenthesis.

*Clones in which the Tc1 TIR was edited were examined for RNA editing in the flanking sequences.

†Clones in which RNA editing was found in internal Tc1 sequences were derived mainly from locus *T10B5*, which contains two repeated Tc1 copies in inverted orientation (one full copy containing nucleotides 1 to 1,611, and one partial copy containing nucleotides 89 to 1,061). Thus, transcripts from this locus can base pair intramolecularly in the region 89 to 1,061. Of the 55 edited nucleotides, 54 are in this region.

Table 2 Detection of read-through transcripts of *Tc1* loci

<i>Tc1</i> locus	Sense	Antisense
<i>R173</i>	1	0
<i>ZK228</i>	4	0
<i>ZK856</i>	9	0
<i>B0213</i>	10	2
<i>T10B5</i>	5	0
<i>C48D1</i>	6	0
<i>Y69A2</i>	3	0
<i>Y37D8A</i>	2	1
<i>Y48G9</i>	1	0
<i>Y48C3A</i>	2	0
<i>Y46G5</i>	0	1
<i>T21B4</i>	2	0
<i>F18C5</i>	1	0
<i>ZK250</i>	26	13
<i>Y39F10A</i>	1	0
<i>ZC334</i>	0	10
<i>T27F6</i>	0	26
<i>F59C6</i>	0	19
13 other loci	0	0
Total no. of clones	71	73

sufficient to encode transposase capable of multiple transposition events.

During RNAi, dsRNA is cleaved by the RNaseIII-like enzyme DCR-1 (ref. 10) to form small interfering RNAs (siRNAs), which are 20–24 nt in length in *C. elegans*¹¹. We detect *Tc1*-specific siRNAs 20–27 nt in length in wild-type worms (Fig. 2a). Compared with the siRNAs that mediate RNAi, the *Tc1* siRNAs have a very low abundance. *Tc1* siRNAs are detected with probes both containing and lacking the TIR, and they are of both sense and antisense polarity (Fig. 2a). This agrees with the finding of three *Tc1*-specific small RNAs (in 3,000 reads) upon random cloning and sequencing of small *C. elegans* RNAs¹²; one corresponds to the *Tc1* TIR, one to sense internal *Tc1* sequences and one to antisense internal *Tc1* sequences. The presence of the internal *Tc1* siRNAs can be explained by the occurrence of dsRNA for internal *Tc1* sequences (for example, from locus *T10B5* that carries a *Tc1* duplication that yields dsRNA for internal *Tc1* sequences; Table 1). Alternatively, these internal *Tc1* siRNAs may represent secondary siRNAs that are generated after RNA amplification by an RNA-directed RNA polymerase (RdRP)¹¹. Strikingly, *Tc1* siRNAs are, like microRNAs and siRNAs induced by RNAi, found in the ribosome-associated fraction¹³.

We analysed whether *Tc1* siRNAs also occur in various mutants. For this, three classes of mutants were tested: (1) a class defective in RNAi only (*rde-1*, *rde-4*)^{3,14}; (2) a class defective in both RNAi and transposon silencing (*mut-7*, *mut-14*, *mut-16*)^{2,15}; and (3) a class defective in transposon silencing only (*pk732*). In all mutants defective in transposon silencing, *Tc1* siRNAs are much less abundant, whereas in mutants defective only in RNAi, siRNAs are readily detected (Fig. 2b). For RDE-1 and RDE-4, this result fits with the their presumed function in detecting and retaining non-endogenous dsRNA¹⁴, although it is unclear how these proteins discriminate between exogenous and endogenous dsRNA. Interestingly, for the mutants defective in both transposon silencing and RNAi, *mut-7* and *mut-16* show an absence of both *Tc1*- and RNAi-specific siRNAs, whereas *mut-14* shows an absence of only *Tc1* siRNAs¹⁶ (for *mut-16*; data not shown). It is possible that MUT-14 (a putative RNA helicase) has a different or an additional role in transposon silencing, such as unwinding dsRNA, that can be fulfilled by DHR-1 during RNAi¹⁴.

As siRNAs drive the sequence-specific RNA degradation that occurs during RNAi¹⁷, transposon siRNAs might mediate the degradation of transposon mRNAs to establish transposon silencing. If this were the case, *Tc1* siRNAs would target recombinant RNAs carrying *Tc1*-derived sequences^{18,19} in a process dependent on genes involved in transposon silencing. To assay for such *trans*-inactivation, transgenic worm lines were generated that carried a

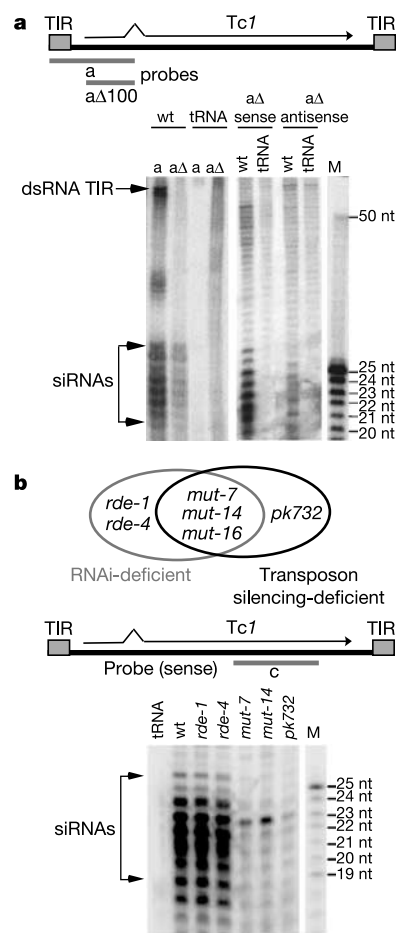


Figure 2 Analysis of *Tc1* siRNAs. **a**, Detection of siRNAs by RNase protection assays on RNA isolated from S100H fractions²⁵ of wild-type worms. Yeast transfer RNA was used as a negative control. Probes containing (probe a) or lacking (probe aΔ100) the TIR region were used to detect siRNAs. siRNAs were also detected using probes b and c indicated in Fig. 1a (data not shown). M, RNA size marker. **b**, Detection of *Tc1* siRNAs by RNase protection assays in RNAi- and/or transposon-silencing-deficient mutants. For each mutant, 100 μg of total RNA was used. In *mut-16*, no *Tc1* siRNAs were detected (results not shown). siRNAs were also detected using probe a indicated in Fig. 1a (data not shown). M, RNA size marker.

chimaeric *gfp* (green fluorescent protein) gene containing the *Tc1* TIR sequence as a carboxy-terminal fusion, transcribed under control of the germline-specific *pie-1* promoter (Fig. 3a). To facilitate detection, GFP was targeted to the nucleus by a histone 2B region. As controls, the *Tc1* TIR region was placed outside the transcribed region, and a 54-nt region of the *unc-22* gene was fused in frame to *gfp* (Fig. 3a). *Trans*-inactivation occurs only in lines that have the *Tc1* TIR region in the mRNA (Fig. 3b) and is dependent on genes required for transposon silencing (*mut-7*, *mut-16* and *pk732*, but not *rde-1*; Fig. 3c). Worms carrying part of the *unc-22* gene in the mRNA express GFP; this expression can be silenced when these animals are fed on *Escherichia coli* expressing dsRNA from *unc-22* (Fig. 3d). When the *Tc1* TIR sequence is present in the non-transcribed region, GFP is expressed in the wild-type background and cannot be silenced upon feeding homologous *Tc1* TIR dsRNA (Fig. 3d). Thus, transgenes that contain the *Tc1* TIR within the *gfp* transcript are specifically silenced, and this silencing is lost in mutants that are defective in transposon silencing.

In *Schizosaccharomyces pombe*, siRNAs have been implicated in guiding transcriptional silencing through the formation of silent chromatin^{20,21}. Therefore, *Tc1* siRNAs may mediate transposon

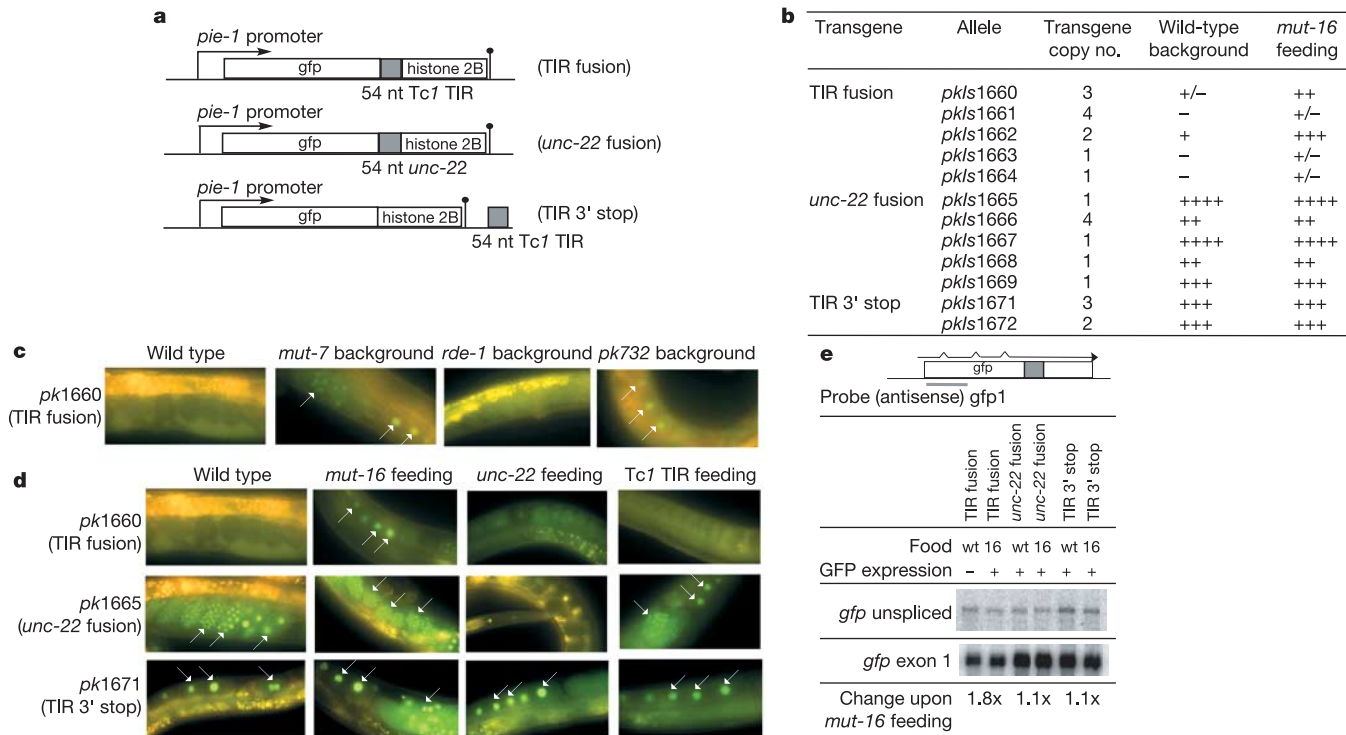


Figure 3 Trans-inactivation of Tc1-containing transgenes. **a**, Schematic representation (not at scale) of chimaeric *gfp* transgenes carrying 54-nt inserts of the Tc1 TIR region or the *unc-22* gene and a histone 2B region for targeting to the nucleus. Transcription start (arrow) and stop (circle) are indicated. Between brackets, the transgene nickname is indicated. **b**, Characteristics of each transgenic line. Transgene copy number was estimated by DNA blot analyses. Strength of GFP expression: +/-, up to 5% of the analysed worms showed weak GFP expression; ++, up to 50% of the analysed worms showed GFP expression; +++, most worms showed GFP expression; +++++, all worms showed very strong GFP expression. Lines that show increased GFP expression upon *mut-16* dsRNA feeding are regarded as silenced. **c**, GFP expression in the germ line of line *pkIs1660*. Animals were analysed in wild-type (N2), *mut-7*(*ne311*), *rde-1*(*ne219*) and *pk732* backgrounds. Arrows indicate nuclei expressing GFP. **d**, Germline GFP

silencing either at the transcriptional level (directing heterochromatin formation) or at the post-transcriptional level (RNA degradation). Our finding that silencing occurs only when the Tc1 TIR is present within the *gfp* mRNA (compare TIR fusion lines with TIR 3' stop lines; Fig 3b), suggests that heterochromatin formation and heterochromatin spreading are not the cause of silencing²² (it should be noted that the Tc1 TIR is 500 bp from the transcription stop in the TIR 3' stop lines). We therefore examined whether there is a post-transcriptional silencing effect. As transposon silencing is restricted to germline tissue, we analysed the germline-specific *gfp*-Tc1 TIR fusion transgene. The silencing of this transgene is relieved by feeding *mut-16* dsRNA, for which two explanations are possible. If post-transcriptional silencing is relieved, the ratio of spliced to unspliced *gfp* mRNA should increase, because post-transcriptional silencing specifically targets cytoplasmic, spliced mRNAs²³. If transcriptional silencing is relieved, the transcription rates should increase and the ratio of spliced to unspliced *gfp* mRNA should remain the same. We find that the ratio of spliced to unspliced *gfp* mRNA increases upon *mut-16* dsRNA feeding specifically in worms carrying the Tc1 TIR fusion transgene (Fig. 3e). This suggests that the silencing observed in the worms on wild-type food occurs at least in part at the post-transcriptional level. Clearly, silencing of the *gfp*-TIR fusion transcript is incomplete; the rather low levels of Tc1 siRNAs combined with the short region of Tc1

expression in worms fed with *E. coli* expressing dsRNA for *mut-16*, *unc-22* or the Tc1 TIR. **e**, Detection of spliced and unspliced *gfp* transcripts by RNase protection assays on total RNA isolated from the lines *pkIs1660*, *pkIs1665* and *pkIs1671* grown on wild-type (OP50) or *mut-16* dsRNA food. The position of the probe in *gfp* is indicated. Pre-treatment with RNaseA showed that the protected fragment marked as unspliced *gfp* transcript is indeed RNA (not DNA). Protected fragments were quantified after generating a line graph, the ratio of spliced versus unspliced mRNA was determined and the change upon *mut-16* dsRNA feeding was calculated. Similar results were obtained when exon 2 of *gfp* was compared with unspliced *gfp* transcript. When *pie-1* and *gfp* mRNA levels were compared by RNA blotting (both genes are under the transcriptional control of the *pie-1* promoter), the ratio of *gfp* to *pie-1* mRNA increases twofold in line *pkIs1660* upon *mut-16* dsRNA feeding (data not shown).

homology (54 nt) in the chimaeric mRNA might account for this. It is intriguing that the silencing machinery of a single transposon can suppress transposition of various transposons that have no sequence similarity² (Tc1, Tc3, Tc5, Tc7). How does the system distinguish between a potentially harmful transposon and a normal gene? Our data suggest that the RNAi machinery recognizes the one common feature of these *C. elegans* transposons: their TIRs, or more specifically the fact that TIR dsRNA is produced when transposons undergo read-through transcription and the transcripts base pair intramolecularly at the TIRs. Even short TIRs (such as the 54-nt Tc1 TIR) make a transposon a target for the natural silencing system, in agreement with the finding that RNAi can be triggered by short (26-bp) stretches of dsRNA²⁴. □

Methods

Strains

The strains and alleles used were *mut7*(*ne311*) (B. Tops, R. Ketting, A. Bathoorn, H. Tabara, C. Mello and R.H.A.P., unpublished), *rde-1*(*ne219*), *rde-4*(*ne299*), *pk732*, *mut-6*(*st702*), *pk739*, *pk724*, *mut-14*(*pk738*), *mut-16*(*pk710*, *pk703*), *fer-1*(*ba576*) and *unc-119*(*dp38*). In crosses, *pk732* was followed using a linked mutation after which homozygosity for the mutator phenotype was confirmed by transposon display analyses.

RNA analyses

RNA isolation, RNA pre-treatment and RNase protection assays were performed as described in refs 7, 11. The probes used for Tc1 were: Tc1-a (1-441), Tc1-b (380-828), Tc1-c (777-1,160), Tc1-d (1,092-1,611), Tc1-TIR (1-54), Tc1-aΔ16 (16-441), Tc1Δ100

(100–441) and Tc1-intron (321–566) (numbers from genomic sequence). The probe used for *gfp* was gfp1 (19–319) (numbered from ATG). Details of the probes used for analyses of Tc3 and Tc5 dsRNA are available on request. 5'-RACE analyses used the SmartII kit (Clontech), SuperScriptII reverse transcriptase (GibcoBRL) and Pwo DNA polymerase. S100H fractions were prepared as described in ref. 25. Standard procedures were used for primer-extension analyses. Sequences of all primers used are available on request. Quantification of protected fragments (RNase protection assays) was performed using ImageQuant software.

Transgenic lines

The chimaeric *gfp* reporter plasmids were produced by inserting various fragments into plasmid pAZ132 (ref. 26). Plasmid pAZ1 (TIR fusion) contained nucleotides 1–54 of Tc1 (genomic sequence) in the sense orientation in the *SgrAI* site. Plasmid pAZ4 (*unc-22* fusion) contained nucleotides 11,137–11,190 of *unc-22* (spliced sequence) in the sense orientation in the *SgrAI* site. Plasmid pAZbb (TIR 3' stop) contained nucleotides 1–54 of Tc1 in the sense orientation in the *BsaBI* site. To prevent transgene silencing due to the presence of high transgene copy numbers²⁷, low-copy-number transgenic lines were generated by ballistic transformation²⁸ using a heptamer adaptor (Bio-Rad). Transformants were generated in *unc-119(dp38)* worms. All lines were selected and they were analysed for GFP expression on both wild-type (OP50) and *mut-16* dsRNA food. Worms were grown under these conditions for two generations. Lines not expressing GFP under any of these conditions were discarded; these included six TIR fusion lines, two *unc-22* fusion lines and ten TIR 3' stop lines. By DNA blot analyses (carried out according to standard procedures), transgene copy number was determined using *SacII*- and *BglII*-digested genomic DNA and *gfp*- and pBlueScript-specific probes. Crosses using *pKIs1660* showed that all three transgene copies in this line reside at one locus and segregate in a mendelian manner. However, the transgenes can be lost (presumably due to recombination), as is apparent from the presence of worms with an *unc-119* phenotype (PCR analyses confirmed transgene loss in these worms). Transgene loss is not uncommon for ballistic-generated transformants and varies for the lines as follows: *pKIs1660*, 1%; *pKIs1661*, 10%; *pKIs1662*, 10%; *pKIs1663*, 2%; *pKIs1664*, 80%; *pKIs1665*, 0%; *pKIs1666*, 1%; *pKIs1667*, 50%; *pKIs1668*, 0%; *pKIs1669*, 0%; *pKIs1671*, 0%; and *pKIs1672*, 0%. Interestingly, transgene loss strongly increases upon crossing to strains defective in transposon silencing (*mut-7* and *pk732* but not *rde-1*); this transgene loss is not dependent on the presence of the Tc1 TIR sequence, as it also occurs upon crossing *mut-7* to *pKIs1665*, an *unc-22* fusion line.

dsRNAs

Plasmids for dsRNA production in *E. coli* comprised pTS302 dsRNA¹¹ (for *unc-22*) and pTS303 (containing nucleotides 1–441 of the Tc1 genomic sequence inserted into the *SmaI* site of vector L4440 (ref. 29)). *E. coli* expressing *mut-16* dsRNA were obtained from well number 17C5 of the *C. elegans* feeding library³⁰.

Received 28 August; accepted 23 September 2003; doi:10.1038/nature02107.

- Emmons, S. W., Yesner, L., Ruan, K. S. & Katzenberg, D. Evidence for a transposon in *Caenorhabditis elegans*. *Cell* **32**, 55–65 (1983).
- Ketting, R. F., Haverkamp, T. H., van Luenen, H. G. & Plasterk, R. H. Mut-7 of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell* **99**, 133–141 (1999).
- Tabara, H. *et al.* The *rde-1* gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99**, 123–132 (1999).
- Fischer, S. E. J., Wienholds, E. & Plasterk, R. H. A. Continuous exchange of sequence information between dispersed Tc1 transposons in the *C. elegans* genome. *Genetics* **164**, 127–134 (2003).
- Vos, J. C., De Baere, I. & Plasterk, R. H. Transposase is the only nematode protein required for *in vitro* transposition of Tc1. *Genes Dev.* **10**, 755–761 (1996).
- Plasterk, R. H. RNA silencing: the genome's immune system. *Science* **296**, 1263–1265 (2002).
- Sijen, T. *et al.* Transcriptional and posttranscriptional gene silencing are mechanistically related. *Curr. Biol.* **11**, 1–20 (2001).
- Aroian, R. V., Field, C., Pruliere, G., Kenyon, C. & Alberts, B. M. Isolation of actin-associated proteins from *Caenorhabditis elegans* oocytes and their localization in the early embryo. *EMBO J.* **16**, 1541–1549 (1997).
- Knight, S. W. & Bass, B. L. The role of RNA editing by ADARs in RNAi. *Mol. Cell* **10**, 809–817 (2002).
- Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
- Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).
- Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13**, 807–818 (2003).
- Caudy, A. A. *et al.* A micrococcal nuclease homologue in RNAi effector complexes. *Nature* **425**, 411–414 (2003).
- Tabara, H., Yigit, E., Siomi, H. & Mello, C. C. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DEXH-box helicase to direct RNAi in *C. elegans*. *Cell* **109**, 861–871 (2002).
- Vastenhouw, N. L. *et al.* A genome-wide screen identifies 27 genes involved in transposon silencing in *C. elegans*. *Curr. Biol.* **13**, 1311–1316 (2003).
- Tijsterman, M., Ketting, R. F., Okihara, K. L., Sijen, T. & Plasterk, R. H. RNA helicase MUT-14-

- dependent gene silencing triggered in *C. elegans* by short antisense RNAs. *Science* **295**, 694–697 (2002).
- Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
- English, J. J., Mueller, E. & Baulcombe, D. C. Suppression of virus accumulation in transgenic plants exhibiting silencing of nuclear genes. *Plant Cell* **8**, 179–188 (1996).
- Sijen, T., Wellink, J., Hiriart, J. B. & Van Kammen, A. RNA-mediated virus resistance: role of repeated transgenes and delineation of targeted regions. *Plant Cell* **8**, 2277–2294 (1996).
- Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
- Reinhart, B. J. & Bartel, D. P. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297**, 1831 (2002).
- Schramke, V. & Allshire, R. Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science* **301**, 1069–1074 (2003).
- Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
- Parrish, S., Fleenor, J., Xu, S., Mello, C. & Fire, A. Functional anatomy of a dsRNA trigger. Differential requirement for the two trigger strands in RNA interference. *Mol. Cell* **6**, 1077–1087 (2000).
- Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
- Praitis, V., Casey, E., Collar, D. & Austin, J. Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. *Genetics* **157**, 1217–1226 (2001).
- Kelly, W. G. & Fire, A. Chromatin silencing and the maintenance of a functional germline in *Caenorhabditis elegans*. *Development* **125**, 2451–2456 (1998).
- Wilm, T., Demel, P., Koop, H. U., Schnabel, H. & Schnabel, R. Ballistic transformation of *Caenorhabditis elegans*. *Gene* **229**, 31–35 (1999).
- Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395**, 854 (1998).
- Fraser, A. G. *et al.* Functional genomic analysis of *C. elegans* chromosome 1 by systematic RNA interference. *Nature* **408**, 325–330 (2000).

Acknowledgements We thank R. Ketting for help in experiments and discussions, and E. Berezikov for help with the ballistic transformations. We acknowledge S. Fischer, N. Vastenhouw, V. Robert, E. Cuppen, R. May and M. Joosten for helpful discussions or for critically reading the manuscript. This work was supported by a VIDI fellowship from the Dutch Science Foundation (NWO) to T.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.H.A.P. (plasterk@niob.knaw.nl).

corrigendum

An NS3 protease inhibitor with antiviral effects in humans infected with hepatitis C virus

Daniel Lamarre, Paul C. Anderson, Murray Bailey, Pierre Beaulieu, Gordon Bolger, Pierre Bonneau, Michael Bös, Dale R. Cameron, Mireille Cartier, Michael G. Cordingley, Anne-Marie Faucher, Nathalie Goudreau, Stephen H. Kawai, George Kukolj, Lisette Lagacé, Steven R. Laplante, Hans Narjes, Marc-André Poupart, Jean Rancourt, Roel E. Sentjens, Roger St George, Bruno Simoneau, Gerhard Steinmann, Diane Thibeault, Youla S. Tsantrizos, Steven M. Weldon, Chan-Loi Yong & Montse Llinàs-Brunet

Nature **426**, 186–189 (2003).

In this Letter, the ‘Competing interests statement’ should be corrected to: ‘The authors declare competing financial interests: R.E.S. was the clinical investigator and received an honorarium from Boehringer Ingelheim. All the other authors are or were employees of Boehringer Ingelheim.’ □

Keeping track, taking stock

Barcoding, lab management systems and other work-saving aids.

Chemical Thesaurus v3.1

Meta-synthesis www.meta-synthesis.com

Word association for chemists

The Chemical Thesaurus is a storehouse of information about chemical species, interactions and reactions that can be downloaded free of charge from the website. The software can be used in a manner similar to that of a thesaurus built into a word processor, in which users can jump from word to word by meaning, as in 'happy/delighted' or 'crude/raw/uncooked'. This thesaurus takes users from chemical species to chemical species via their chemical reactions, for example: 'natural gas/methane/hydrogen cyanide/adiponitrile/1,6-diaminohexane/nylon 6,6'. It uses a relational database to store information about any chemical reaction process. The software is also extensively hyperlinked. Versions are available for Windows 96, 98, Me, NT4, 2000 and XP, and Mac OS 8.6–10.2.

Xylene-resistant labels

Computer Imprintable Label Systems

www.cils-labels.com

Don't come unstuck in the lab

Computer Imprintable Label Systems (CILS) has extended its range of products to include laboratory labels formulated to withstand extended exposures to xylene, IPA and other harsh laboratory chemicals. Specially formulated high-specification adhesives and coatings protect the computer-printed data and keep the labels adhered to the slides during chemical exposure. They can be supplied in any design and size and are compatible with standard laboratory thermal transfer printers for in-laboratory printing of variable computer-generated text, barcodes and other customized symbols. Applications include cytopathology, histology, virology, clinical chemistry and labelling of tissue and specimen slides.



More than just barcodes, from CILS.

BP1344 transfer printer

Triple Red

www.triplered.com

Impressive output

Triple Red's compact desktop thermal transfer printer produces high-definition printed labels for cryovials, Eppendorfs, microtitre plates and microscope slides. A resolution of 203 d.p.i. provides clear and durable images including text, graphics and barcodes. The system can be linked to a PC or operated in a stand-alone mode via an external keyboard. Optional links to robotic systems are also available. The machine is suitable for high-volume printing of labels using an existing database or spreadsheet. Labels suitable for samples used in extreme temperatures (as low as -152°C), such as those found in cryogenics, can also be produced. Other features of the BP1344 include a facility to design bespoke labels and seamless links with laboratory information management systems.

VirtualLab

Mettler Toledo AutoChem

www.mtautochem.com

Speed up R&D

VirtualLab is an integrated informatics platform for workflow automation, instrument integration and electronic laboratory notebooks. It can be used to integrate and automate information-rich R&D workflows and to assemble all the elements of the experimental process into a single software environment. Experimental design, execution and analysis can be performed in a way that is similar to a researcher's natural manner of working. Common information need only be entered once, complex unit conversions are performed automatically, and design parameters are easily applied to generate multiple experiments as needed. VirtualLab then collects and integrates process and analytical data from multiple laboratory instruments as well as traditional bench chemistry activities. The resulting data for analysis are stored in a structured, portable and searchable format.

Sani-Tech STHT-R-HD

Saint-Gobain Performance Plastics

www.biopharm.saint-gobain.com

Heavy-duty silicone hose

This high-purity grade silicone hose is designed for elevated pressure applications where strength, durability and continuous length are critical. The hose is constructed of biopharmaceutical grade platinum silicone, is ultra-flexible, with improved bend radius over wire-reinforced tubing, and is available in



Four colour image: Saint-Gobain's new hose.

custom colours for line identification and long lengths to reduce installation time. The hoses are fully autoclavable and sterilizable, impart no taste or odour, and meet all USP Class VI, FDA, ISO 10993 and EP 3.1.9 standards.

Limited-volume polypropylene shell vial

J. G. Finneran Associates www.jgfinneran.com

Form follows function

The 12 mm × 32 mm limited volume shell vial in polypropylene from J. G. Finneran features a conical bottom to facilitate maximum sample evacuation. The polypropylene material allows for easy disposal. When used with the company's assortment of 12 mm conical coloured plugs, any experiment can be colour-coded. The 1.5 ml vial can be accommodated in the same equipment that uses glass or plastic flat bottom vials.

Automated syringes

Socorex

www.socorex.ch

Take the plunge

Socorex's range of automatic dispensing syringes includes models with dosing capabilities from 0.025 ml to 20 ml. All syringes have a chrome-plated body, interchangeable



Get a grip: two models of syringe from Socorex.

glass barrel, volume-setting mechanism with locking nut, two-way valve system and spring-loaded plunger, allowing automatic filling of each dose. Tightness is achieved without an O-ring through optimal matching of the plunger with the ground glass barrel. The syringes are autoclavable at 121 °C for 20 minutes, thus providing added safety in sterile applications such as distribution of hot agar and culture media, filtration or sampling. Pistol or two-ring handle models made of materials with superior corrosion are available, as well as basic versions suitable for aqueous liquids.

CISPro CIS enhancement

ChemSW www.chemsw.com

You've got mail

ChemSW has added a customizable, automatic e-mail notification feature to its CISPro Chemical Inventory System (CIS). Users can establish criteria for the automatic generation of alerts from the inventory system. The software program will monitor chemical inventory and issue reports when re-orders are needed or when chemicals on hand are reaching maximum storage levels, as determined by regulations that are built in to the system. It will also monitor expiry dates of stored chemicals and

alert users when chemicals are about to expire, thereby minimizing disposal costs. Individual users can determine the circumstances under which they want reports generated, specify custom reports, and designate who should receive reports via email. Anyone in a company can receive the alerts, including those who are not direct system users.

CyBi-SIENA

CyBio www.cybio-ag.com

Check your plates

CyBi-SIENA is a software program for quality control and correction of systematic errors in high-throughput screening data sets. This algorithmic solution detects plate local effects as well as trends within a sequence of microplates, with no user interaction. Recognized patterns are automatically compensated for. An integrated plate inspector enables users to compare the original data with the corrected data, and visualize the correction. Processing time for data correction per plate lies within the range of a few seconds.

Low-profile microplates

Genetix www.genetix.com

These stack up well

The low height profile (10 mm) of these microplates from Genetix increases storage capacity by up to a third, saving valuable freezer space. They are fully compatible with most liquid-handling equipment and are thus particularly suitable for high-throughput use. Each plate has alpha-numeric grid referencing for easy location of wells, and cut-away corners for correct lid orientation. This aids stacking stability and helps to minimize cross-contamination between wells. The plates are made of highest-grade polystyrene to ensure near-zero leachates and reduced cell and tissue binding to the plate surface. They are supplied sterile in a variety of pack and case sizes, with Genetix plate lids included as standard.

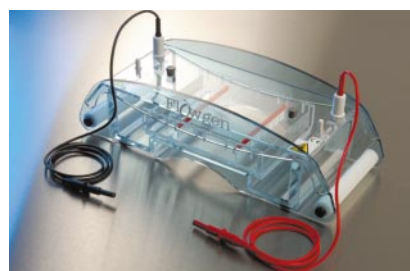
Horizontal electrophoresis tanks

Flowgen www.flowgen.co.uk

Smooth operators

Flowgen's range of seven horizontal electrophoresis tanks feature a curved design with fewer joints for minimal leakage. Operator visibility is improved and the smooth design with few 'nooks and crannies' means the units are easier to clean. Access to the gel tray is also improved. The electrodes are embedded in the side wall of the tank, reducing damage to the vertical section of PTFE sheathed wire and prolonging the life of the tank. A fitted integral locking lid system ensures that the lid will not move once in position, eliminating the potential for electrodes to be snapped off. Ventilation holes at each end of the unit allow air circulation, reducing evaporation and band streaking and preventing the unit from over-

heating. Curved handles mean that users will not catch the tank on other objects when moving it about in the laboratory.



One in seven: from Flowgen's new range.

Auto HT-12

Genevac www.genevac.co.uk

Robotic evaporator

Genevac's robotic centrifugal evaporator is designed to help accelerate compound purification, plate re-formatting and plate replication. The Auto HT-12 is a robot-ready version of the HT-12 large-scale solvent evaporator. It incorporates remote access doors and an indexing rotor so that sample holders always line up with the robot arm. In addition, full remote control and data logging via the built-in RS232C interface make it easy to integrate the unit with most laboratory automation systems. Control software automatically detects when samples are dry and shuts the unit down, safeguarding samples from thermal degradation. A triple-deck rotor can accommodate 48 shallow-well plates simultaneously. The unit has a small footprint and can be positioned in any standard fume hood. It is especially suitable for drying large numbers of samples of broadly similar mass and format, such as microplates and HPLC autosampler vials.

GeneTools MADGE

Syngene www.syngene.com

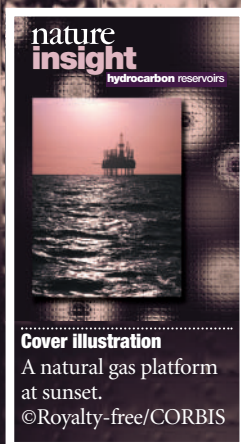
Set a track record

GeneTools MADGE is a robot-compatible software program for high-throughput gel studies. Microtitre array diagonal gel electrophoresis (MADGE) gels are run in a grid format based on a microtitre plate and have the potential for up to 768 tracks within a single gel. Users can select the degree of rotation and electrophoresis direction as well as the number of 96-well data sets using columns and rows. The grid allows manual realignment of track positions, making track location quick and simple. Marker tracks (standards) can then be selected and their peaks assigned, enabling calculation of molecular weight for all unknown peaks in the gel. Results are exportable to Microsoft Excel, and reports can be generated within the package or via Microsoft Word.

These notes are compiled in the Nature office from information provided by the manufacturers.

nature insight

hydrocarbon reservoirs



The importance of fossil fuels to human society cannot be overstated. Naturally formed reservoirs of hydrocarbons occur in a variety of geological contexts (most notably as oil and gas) and are exploited to satisfy the majority of our energy needs. Such resources are finite, yet the demand for fossil fuels is growing as the industrialization of the world continues apace. The usage of fossil fuels also comes at a cost, for example, they are strongly implicated as the main driver of climate change. Consequently, the societal impact of hydrocarbons is multi-faceted, encompassing economics, politics and the environment — all issues that are the subject of ongoing and often heated public debate.

But until we find economically viable alternative sources of energy to support our energy-hungry way of life, research is needed to further our understanding of hydrocarbon- and reservoir-formation processes, in order to exploit diminishing supplies.

It has long been appreciated that a rethink is required concerning both the location and nature of the reserves that have yet to be exploited: the 'low-hanging apples' have long since been plucked, and greater ingenuity is required to both locate and economically extract the deposits of hydrocarbons that remain. And as the industry extends its reach into increasingly hostile environments (such as deep water) — or turns its attention to unconventional classes of hydrocarbon that have until now been largely ignored (for example, gas hydrates) — it finds itself faced by a host of unfamiliar scientific challenges.

Despite the maturity of the industry as a whole, its academic needs have never been greater. No longer is it simply a case of seek-and-drill and wait for the fuel (and profits!) to flow: the modern hydrocarbon engineer increasingly needs to draw on a wealth of scientific expertise, integrating disciplines as diverse as physics, chemistry and biology to both understand and successfully exploit previously intractable reservoirs. This Insight aims to give a flavour of these diverse scientific challenges and, in so doing, highlights the need for a 'renaissance' approach to industrial hydrocarbon research.

Karl Ziemelis Physical Sciences Editor

Web focus: Fossil fuels To accompany this Insight, *Nature* presents a special web focus on the implications of hydrocarbon fuel usage, including collections of news, features and research.

318 commentary:
Hydrocarbons and the evolution of human culture
C. Hall, P. Tharakan, J. Hallock, C. Cleveland & M. Jefferson

323 overview:
The long-term carbon cycle, fossil fuels and atmospheric composition
R. A. Berner

327 review articles:
Organic-inorganic interactions in petroleum-producing sedimentary basins
J. S. Seewald

334 Understanding the thermal evolution of deep-water continental margins
N. White, M. Thompson & T. Barwise

344 Biological activity in the deep subsurface and the origin of heavy oil
I. M. Head, D. M. Jones & S. R. Larter

353 Fundamental principles and applications of natural gas hydrates
E. D. Sloan Jr

360 commentary:
Integration of geoscience and engineering in the oil industry — just a dream?
B. A. Stankiewicz

Editor, *Nature*: Philip Campbell
Publisher: Peter Collins
Insights Editor: Karl Ziemelis
Production Editor: Maria Hodges

Art Editor: Martin Harrison
Layouts: Nadia Rooney
Production: Sue Gray
Sponsorship: Mark Greene

Diagrams: Ann Thomson
Mick Ayres
Glyn Watson

Hydrocarbons and the evolution of human culture

Charles Hall^{1,2}, Pradeep Tharakan^{1,3}, John Hallock¹, Cutler Cleveland^{4,5} and Michael Jefferson⁶

¹Departments of Environmental and Forest Biology and Graduate Program in Environmental Sciences, State University of New York College of Environmental Science and Forestry, Syracuse, New York 13210, USA (SUNY ESF); ²Flathead Lake Biological Station, University of Montana, Polson, Montana, USA; ³Maxwell School of Citizenship and Public Affairs, Syracuse University, Syracuse, New York, USA; ⁴Center for Energy and Environmental Studies, Boston University, 675 Commonwealth Avenue, Boston, Massachusetts 02215, USA; ⁵Department of Geography at Boston University, 675 Commonwealth Avenue, Boston, Massachusetts 02215, USA; ⁶Global Energy & Environmental Consultants, The Old Stables, Felmersham, Bedfordshire MK43 7HJ, UK (e-mails: chall@esf.edu; cutler@bu.edu; jeffers@dircon.co.uk)

Most of the progress in human culture has required the exploitation of energy resources. About 100 years ago, the major source of energy shifted from recent solar to fossil hydrocarbons, including liquid and gaseous petroleum. Technology has generally led to a greater use of hydrocarbon fuels for most human activities, making civilization vulnerable to decreases in supply. At this time our knowledge is not sufficient for us to choose between the different estimates of, for example, resources of conventional oil.

The history of human culture can be viewed as the progressive development of new energy sources and their associated conversion technologies. These developments have increased the comfort, longevity and affluence of humans, as well as their numbers. Most of these energy technologies rely on chemical bonds of hydrocarbons. Nature has favoured the storage of solar energy in the hydrocarbon bonds of plants and animals, and human cultural evolution has exploited this hydrocarbon energy profitably.

A key event in the evolution of human society was the development of spear heads and knife blades, devices that allowed humans to exploit a much broader and larger animal-resource base for food and skins. Another was the harnessing of the energy in the hydrocarbon bonds of wood, using fire, which allowed humans to exploit even more food resources, to smelt metals and to bake ceramics. All these developments assisted humans in their exploitation of colder, more northerly ecosystems. The most important of these new energy-based technologies was agriculture, which redirected photosynthetic energy from natural to human food chains.

The principal energy sources of antiquity were all derived directly from the sun: human and animal muscle power, wood, flowing water and wind. About 300 years ago, the industrial revolution began with stationary wind-powered and water-powered technologies, which were essentially replaced by fossil hydrocarbons: coal in the nineteenth century, oil since the twentieth century, and now, increasingly, natural gas. The global use of hydrocarbons for fuel by humans has increased nearly 800-fold since 1750 and about 12-fold in the twentieth century.

Hydrocarbon-based energy is important for the three main areas of human development: economic, social and environmental¹ (Fig. 1). Both the popular and some scientific presses have suggested that we have entered a 'post-industrial' society, where computers and, more generally, human knowledge have replaced raw energy and materials in the generation of wealth. 'Bottom-up analysis', applied by engineers, physicists and some environmentalists, suggests that a substantial decoupling of energy and economic production is now underway². Nevertheless, there continues to be a strong connection between energy and economic activity for most industrialized³ and developing economies⁴.

Top-down macroeconomic analysis indicates that where there is a decline in the ratio of energy to gross domestic product in industrial nations, it is due principally to a shift to higher quality fuels, improvements in fuel efficiency driven by higher fossil fuel prices and structural changes in national economies⁵. Energy prices have an important effect on almost every major aspect of macroeconomic performance, because energy is used directly and indirectly in the production of all goods and services. Both theoretical models and empirical analyses of economic growth suggest that a decrease in the rate of increase in energy availability will have serious impacts⁶. For example, most US recessions after the second World War were preceded by rising oil prices, and there tends to be a negative correlation between oil price changes and both stock prices and returns⁷ in countries that are net importers of oil and gas. Energy prices have also been key determinants of inflation and unemployment.

There is a strong correlation between per capita energy use and social indicators such as the UN's Human Development Index. By contrast, the use of hydrocarbons to meet economic and social needs is a major driver of our most important environmental changes, including global climate change, acid deposition, urban smog and the release of many toxic materials. Increased access to energy also provided the means to deplete or destroy once-rich resource bases, from the megafaunal extinctions associated with each new invasion of spear-equipped humans, to the destruction of natural ecosystems and soils through, for example, over-fishing and intensive agriculture and other types of development. Such problems are exacerbated by the increase in human populations that each new technology has allowed, as well as the overdependence of societies on those once-abundant resources. Energy is a double-edged sword.

How long can we depend on oil?

At present, oil supplies about 40% (natural gas 25%) of the world's non-solar energy, and most future assessments indicate that the demand for oil will increase substantially. What do we know about the future of oil? Predictions of impending oil shortages are as old as the industry itself, and the literature is full of arguments between 'optimists' and 'pessimists' about how much oil there is and what other resources might be available. There are four principal issues



Figure 1 Former oil fields in Southern Louisiana. This area was once a productive salt marsh that contributed greatly to Louisiana's extremely rich fisheries. The numerous straight lines in the picture are the result of past dredging operations to float oil 'rigs' into areas where they extracted oil. The dredging has resulted in enormous land erosion, exacerbated by sea-level increase, which in turn was partly a consequence of greenhouse gas emissions from burning that oil. The oil production is gone, but the environmental degradation remains. Now oil operations have moved offshore, where there are 4,000 energy-intensive oil platforms off Louisiana. There are still oil operations in the coastal region that receive offshore and foreign oil. Louisiana, once the fourth largest oil-producing state (and number one for gas) in the USA, now uses as much oil as it produces and serves mainly as an energy-intensive conduit and processor for foreign oil moving into the US interior (courtesy of Louisiana Department of Wildlife and Fisheries).

that we need to understand to assess the availability of oil, and, by extension, other hydrocarbons, for the future. We need to know: first, the quality of the reserves; second, the quantity of the reserve; third, the likely patterns of exploitation of the resource over time; and fourth, who gets, and who benefits from, the oil. All of these factors ultimately affect the economics of oil production and use.

Quality of petroleum

What we call oil is actually a large family of diverse hydrocarbons whose physical and chemical qualities reflect the different origins and, especially, different degrees of natural processing of these hydrocarbons⁸. In general, humans have exploited the large reservoirs of shorter-chain 'light' oil resources first because larger reservoirs are easier to find and exploit, and lighter oils are more valuable and require less energy to extract and refine. Therefore, over time in mature regions, lower quality has often required the exploitation of increasingly small, deep, offshore and heavy resources (Figs 2 and 3). Progressive depletion also means that oil in older fields that once came to the surface through natural drive mechanisms, such as gas pressure, must now be extracted using energy-intensive secondary and enhanced technologies. Thus, technological progress is in a race with the depletion of higher-quality resources. Another aspect of the quality of an oil resource is that oil reserves are normally defined by their degree of certainty and their ease of extraction, classed as 'proven', 'probable', 'possible' or 'speculative'. In addition, there are unconventional resources such as heavy oil, deep-water oil, oil sands and shale oils that are very energy intensive to exploit.

Quantity of petroleum

Most estimates of the quantity of conventional oil resources remaining are based on 'expert opinion', which is the carefully considered opinion of geologists and others familiar with a particular region (Table 1). The ultimate recoverable resource (URR) is the total quantity of oil that will ever be produced, including the nearly 1 trillion barrels extracted to date. Recent estimates of URR for the world have tended to fall into two camps. Lower estimates come from several high-profile analysts

with long histories in the oil industry. They suggest that the URR is no greater than about 2.3 trillion barrels, and may even be less (for example, ref. 9). A higher estimate of 3 trillion barrels is the middle estimate, and 4 trillion is the highest estimate, from the most recent study by the US Geological Survey (USGS)^{10,11}. About half of the roughly 1.4 trillion barrels that the USGS predicts remain to be discovered are from new discoveries and about half are from reserve growth. The latter describes the process by which technical improvements and correction of earlier conservative estimates increase the projected recovery from existing fields. This relatively new addition to the USGS methodology is based on experience in the US and a few other well-documented regions. The new totals assume, essentially, that petroleum reserves everywhere in the world will be developed with the same level of technology, economic incentives and efficacy as in the US. Time will tell the extent to which these assumptions are realized.

Pattern of use over time

The best-known model of oil production was proposed by Marion King Hubbert, who proposed that the discovery, and production, of petroleum over time would follow a single-peaked, symmetric bell-shaped curve with a peak in production when 50% of the URR had been extracted. This hypothesis seems to have been based principally on Hubbert's intuition, and it was not a bad guess as he famously predicted in 1956 that US oil production would peak in 1970, which in fact it did¹². Hubbert also predicted that the US production of natural gas would peak in about 1980, which it did, although it has since shown signs of recovery. He also predicted that world oil production would peak in about 2000. There was a slight downturn in world production in 2000, but production in the first half of 2003 is running slightly above the rate in 2000.

In the past decade, a number of 'neohubbertarians' have made predictions about the timing of peak global production using several variations of Hubbert's approach. Various forecasts of the year of the global peak have ranged from one predicted for 1989 (made in 1989) to many predicted for the first decade of the twenty-first century to one as late as 2030 (ref. 9). Their predictions begin with an a priori

assumption about the volume of ultimately recoverable oil. Most of these studies assumed world URR volumes of roughly 2 trillion barrels and that oil production would peak when 50% of the ultimate resource had been extracted. In comparison, the USGS low estimate (which they state has a 95% probability of being exceeded) is 2.3 trillion barrels. One analysis fitted the left-hand side of Hubbert-type curves to data on actual production while constraining the total quantity under the curve to 2, 3 and 4 trillion barrels for world URR. The resultant peaks were predicted to occur from 2004 to 2030.

Other forecasts for world oil production do not rely on such curve-fitting techniques to make future projections and/or a priori assumptions about URR. According to the most recent forecast by the US Energy Information Agency (EIA) (2003), world oil supply in 2025 will exceed the 2001 level by 53% (ref. 13). The EIA reviewed five other world oil models and found that all of them predict that production will increase in the next two decades to around 100 million barrels per day, substantially more than the 77 million barrels per day

produced in 2001. Several of these models rely on the new USGS estimates of URR for oil. It should be noted that almost all oil-supply forecasts for which we are able to examine the predictions against reality had a dismal track record, regardless of method. Most recent results of curve-fitting methods showed a consistent tendency to predict a peak within a few years, and then a decline, no matter when the predictions were made¹⁴. It is now a well-established fact that economic and institutional factors, as well as geology, were responsible for the US peak in production in 1970 (ref. 15), forces that are explicitly excluded from the curve-fitting models. Thus, the ability (or the luck) of Hubbert's model (and its variants) to forecast production in the 48 lower states (that is, contiguous) accurately cannot necessarily be extrapolated to other regions. It is too early to tell.

Economic forecasts fare no better in explaining US oil production in the lower 48 states. In the period after the Second World War, oil production often increased as oil prices decreased, and *viva versa*¹⁶, a behaviour that is exactly the opposite of predictions of economic theory. Economic theory also assumes that oil prices will follow an 'optimal' path towards the choke price — the price at which demand for oil falls to zero and the market signals a seamless transition to substitutes. In fact, even if such a path exists, prices may not increase smoothly because empirical evidence indicates that producers respond differently to price increases than they do to price decreases¹⁵. Significant deviation from basic economic theory undermines the *de facto* policy for managing the depletion of conventional oil supplies — a belief that the competitive market will generate a smooth transition from oil. It also suggests the need for a greater degree of government intervention in the transition from oil than is currently envisaged by most policy makers.

Geography

Oil is used by all of the ~220 nations of the world, but significant amounts are produced by only about 42 countries, 38 of which export important amounts. This number is likely to change because of the depletion of the once-vast resources of North and South America, and owing to the increasing domestic use of oil by many of the exporters. The number of exporters outside the Middle East and the former Soviet Union will drop in the coming decades, perhaps sharply, which in turn will greatly reduce the supply diversity to the 180 or so importing nations¹⁷. Such an increase in reliance on West African, former Soviet Union and especially Persian Gulf oil has many strategic, economic and political implications.

Energy and political costs of getting oil

The future of oil supplies is normally analysed in economic terms. But the economic terms are likely to be dependent on other costs. In earlier work we summarized the energy costs of obtaining US oil and other energy resources and found, in general, that the energy returned on energy invested (EROI) tended to decline over time for all energy resources examined. This includes the energy cost of obtaining oil by trading (energy-requiring) goods and services for energy itself¹⁸. For example, the EROI of oil in the US has decreased from a value of at least 100 to 1 for oil discoveries in the 1930s, to about 17 to 1 today for oil and gas extraction (Fig. 4). We are not aware of such estimates for other parts of the world, although we do know that both heavy oil in Venezuela and tar sands in Alberta require a very large part of the energy produced as well as substantial supplies of hydrogen from natural gas to make the oil fluid. The very low economic cost of finding or producing new oil supplies in the Arabian Peninsula implies that it has a very high EROI value, which in turn supports the probability that productivity will be concentrated there in future decades. Alternative liquid fuels such as ethanol from corn have a very low EROI. An EROI of much greater than 1 to 1 is needed to run a society, because energy is also required to make the machines that use the energy, feed, house, train and provide health care for necessary workers and so on.

No one who watches the news can fail to be aware of the impor-



Figure 2 Oil tanker from an oil rig in the North Sea. Copyright Shell International Photographics Services.



Figure 3 Deck of a North Sea oil well. Copyright Shell International Photographics Services.

tance of cultural and political differences between those nations that have the most oil and those that import it. How these factors will play out over the next few decades is extremely important but also impossible to predict. Most of the remaining oil reserves are in Southern Russia, the Middle East and North and West Africa, countries or regions with either Muslim governments or significant Muslim groups. For a long period, frustration and resentment has been building up among Muslim populations, not least because of their perception that the main Western powers have failed to generate even-handed policies to address the conflict in the Middle East over the past half-century. It also is the case that the huge revenues earned by the oil-exporting nations have been very unevenly distributed among their respective populations, adding to internal and external pressure to adopt a more equitable approach to human development. Suffice it to say that there will continue to be high risks of international and national terrorism, overthrow of existing governments and deliberate supply disruption in the years ahead. In addition, exporting nations may wish to keep their oil in the ground to maintain their target price range. Thus, there are considerable political and social uncertainties that could result in less oil being available than existing models predict.

Our need to reduce supply uncertainties

Many once-proud ancient cultures have collapsed, in part, because of their inability to maintain energy resources and societal complexity¹⁹. Our own civilization has become heavily dependent on enormous flows of cheap hydrocarbons, partly to compensate for other depleted resources (for example, fertilizers and long-range fishing boats), so it seems important to assess our main energy alternatives. Some of our most promising new oil fields have turned out to be very disappointing²⁰. If indeed we are approaching the oil scarcity that some predict, it is not reflected in price and few investments are being made at the scale required. An even greater problem may be that an increasing number of decision-makers sense that the market has resolved this issue before and will and should do so again, and also that government programmes are too inefficient to resolve possible

Table 1 Published estimates of world oil ultimate recovery

Source	Volume (trillions of barrels)
USGS, 2000 (high) (ref. 11)	3.9
USGS, 2000 (mean) (ref. 11)	3.0
USGS, 2000 (low) (ref. 11)	2.25
Campbell, 1995	1.85
Masters, 1994	2.3
Campbell, 1992	1.7
Bookout, 1989	2.0
Masters, 1987	1.8
Martin, 1984	1.7
Nehring, 1982	2.9
Halbouty, 1981	2.25
Meyerhoff, 1979	2.2
Nehring, 1978	2.0
Nelson, 1977	2.0
Folinsbee, 1976	1.85
Adam and Kirby, 1975	2.0
Linden, 1973	2.9
Moody, 1972	1.9
Moody, 1970	1.85
Shell, 1968	1.85
Weeks, 1959	2.0
MacNaughton, 1953	1.0
Weeks, 1948	0.6
Pratt, 1942	0.6

Source: ref. 21.

impending energy problems. We view this as a recipe for disaster, and it is enhanced by the failure of science to be used as fully as it should be. Thus, in 2003, the state of oil-supply modelling is in some ways no different than it was in Hubbert's time; in other words, a wide range of opinion exists.

What can science do to help resolve this uncertainty? Our principal conclusion is that these critical issues could be and should be the province of open scientific analysis in visible meetings where 'all sides' attend and argue. This analysis should be informed by the peer-review process, statistical analysis, hypothesis-generating and testing, and so on, rather than by the experts one chooses. These issues should be the basis of open competitive government grant programmes, graduate seminars and even undergraduate courses in universities, and our courses in economics should become at least as



Figure 4 Service vessels and infrastructure for off-shore oil facilities in Southern Louisiana. (C. Hall)

much about real biophysical resources, such as hydrocarbon reserves, as about market mechanisms. And we need to think much harder about the alternatives.

The future: other technologies

The world is not about to run out of hydrocarbons, and perhaps it is not going to run out of oil from unconventional sources any time soon. What will be difficult to obtain is cheap petroleum, because what is left is an enormous amount of low-grade hydrocarbons, which are likely to be much more expensive financially, energetically, politically and especially environmentally. As conventional oil becomes less important, society has a great opportunity to make investments in a different source of energy, one freeing us for the first time from our dependence on hydrocarbons.

There are a wide range of options and an equally wide range of opinions on the feasibility and desirability of each. Nuclear power faces formidable obstacles. Experience of the past several decades has shown that electricity from nuclear power plants is an expensive form of power when all public and private costs are considered. Nuclear power generates high-level radioactive wastes that remain hazardous for thousands of years and increases the likelihood of nuclear-weapon proliferation. These are high costs to impose on future generations. Even with improved reactor design, the safety of nuclear plants remains an important concern. Can these technological, economic, environmental and public safety problems be overcome? This remains an unanswered question.

Renewable energies present a mixed bag of opportunities. Some have clear advantages over hydrocarbons in terms of economic viability, reliability, equitable access and environmental benefits. In favourable locations, wind power has a high EROI. The cost of photovoltaic (solar electric) power has come down sharply, making it a viable alternative in areas without access to electricity grids. With proper attention to environmental concerns, biomass-based energy generation is competitive in some cases relative to conventional hydrocarbon-based energy generation. By contrast, liquid-fuel production from grain and solar thermal power has a relatively low EROI. Hydrogen is an energy carrier, not an energy source, but energy and environmental communities have shown enormous interest in its potential. Hydrogen generated from renewable energy sources or electricity-driven hydrolysis is currently expensive for most applications, but it merits further research and development.

Subsidies and externalities, social as well as environmental, affect energy markets. With few exceptions, these subsidies and externalities tilt the playing field towards conventional sources of energy. This presents a clear case for public-policy intervention that would encourage the research, development and adoption of renewable

forms of energy. Policy intervention, in concert with ongoing private investment, will speed up the process of sorting the wheat from the chaff in the portfolio of feasible renewable energy technologies. It is time to think about possibilities other than the next cheapest hydrocarbons, if for no other reason than to protect our atmosphere, and for this task we must use all of our science, both natural science and social science, more intelligently than we have done so far. □

doi:10.1038/nature02130

1. Munasinghe, M. The sustainomics trans-disciplinary meta-framework for making development more sustainable: applications to energy issues. *Int. J. Sustain. Dev.* **5**, 125–182 (2002).
2. Interlaboratory Working Group. *Scenarios for a Clean Energy Future* <http://www.ornl.gov/ORNL/Energy_Eff/CEE.htm> (Lawrence Berkeley National Laboratory LBNL-44029, Berkeley, California, 2000).
3. Hall, C. A. S., Lindenberger, D., Kummel, R., Kroeger, T. & Eichhorn, W. The need to reintegrate the natural sciences with economics. *BioScience* **51**, 663–673 (2001).
4. Tharakan, P. J., Kroeger, T. & Hall, C. A. S. Twenty-five years of industrial development: a study of resource use rates and macro-efficiency indicators for five Asian countries. *Environ. Sci. Policy* **4**, 319–332 (2001).
5. Kaufmann, R. K. The mechanisms for autonomous increases in energy efficiency: a cointegration analysis of the US energy/GDP ratio. *Energy J.* (in the press).
6. Smulders, S. & de Noij, M. The impact of energy conservation on technology and economic growth. *Resource Energy Econ.* **25**, 59–79 (2003).
7. Sadosky, P. Oil price shocks and stock market activity. *Energy Econ.* **21**, 449–469 (1999).
8. Tissot, B. P. & Welt, D. H. *Petroleum Formation and Occurrence* (Springer-Verlag, New York, 1978).
9. Campbell, C. J. & Laherrère, J. H. The end of cheap oil. *Sci. Am.* **278**, 78–83 (1998).
10. United States Geological Survey (USGS) *The World Petroleum Assessment 2000* <www.usgs.gov> (2003).
11. United States Geological Survey (USGS) *United States Department of Long Term World Oil Supply* <http://www.eia.doe.gov/pub/oil_gas/petroleum/presentations/2000/long_term_supply/index.htm> (2000).
12. Hubbert, M. K. *Energy Resources (Report to the Committee on Natural Resources)* (National Academy of Sciences, Washington DC, 1962).
13. Energy Information Administration, US Department of Energy. *International Outlook 2003*. Report No. DOE/EIA-0484 (2003), Table 16 at <<http://www.eia.doe.gov/oiaf/ieo/oil.html>> (2003).
14. Lynch, M. C. Forecasting oil supply: theory and practice. *Q. Rev. Econ. Finance* **42**, 373–389 (2002).
15. Kaufmann, R. K. & Cleveland, C. J. Oil Production in the lower 48 states: economic, geological and institutional determinants. *Energy J.* **22**, 27–49 (2001).
16. Kaufmann, R. K. Oil production in the lower 48 states: Reconciling curve fitting and econometric models. *Res. Energy* **13**, 111–127 (1991).
17. Hallock, J., Tharkan, P., Hall, C., Jefferson, M. & Wu, W. Forecasting the availability and diversity of geography oil supplies. *Energy* (in the press).
18. Cleveland, C. J., Costanza, R., Hall, C. A. S. & Kaufmann, R. Energy and the United States economy: a biophysical perspective. *Science* **225**, 890–897 (1984).
19. Tainter, J. *The Collapse of Complex Systems* (Cambridge Univ. Press, Cambridge, 1988).
20. Cooper, C. & Pope, H. Dry wells belie hope for big Caspian reserves. *Wall Street J.* 12 October (1998).
21. Hakes, J. *Long Term World Oil Supply: a Presentation Made to the American Association of Petroleum Geochemists, New Orleans, Louisiana* <http://www.eia.doe.gov/pub/oil_gas/petroleum/presentations/2000/long_term_supply/index.htm> (2000).

Supplementary Information accompanies the paper at www.nature.com/nature.

Acknowledgements We thank S. Ulgiati, R. Kaufmann and C. Levitan for discussions, J. Laherrère for insights into the functioning of the oil industry and for directing us to the recent ASPO data.

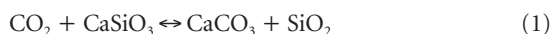
The long-term carbon cycle, fossil fuels and atmospheric composition

Robert A. Berner

Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06520-8109, USA (e-mail: robert.berner@yale.edu)

The long-term carbon cycle operates over millions of years and involves the exchange of carbon between rocks and the Earth's surface. There are many complex feedback pathways between carbon burial, nutrient cycling, atmospheric carbon dioxide and oxygen, and climate. New calculations of carbon fluxes during the Phanerozoic eon (the past 550 million years) illustrate how the long-term carbon cycle has affected the burial of organic matter and fossil-fuel formation, as well as the evolution of atmospheric composition.

The carbon cycle is a variety of processes that take place over timescales ranging from hours to millions of years. Processes occurring over shorter periods include photosynthesis, respiration, air–sea exchange of carbon dioxide and humus accumulation in soils. However, it is the long-term carbon cycle, occurring over millions of years, that is of interest when considering the origin of fossil fuels. The long-term cycle, shown in Fig. 1, is distinguished by the exchange of carbon between rocks and the surficial system, which consists of the ocean, atmosphere, biosphere and soils. The long-term carbon cycle is the main controller of the concentration of atmospheric carbon dioxide and (along with the sulphur cycle) atmospheric oxygen over a geological timescale^{1,2}. It can be represented succinctly by the generalized reactions^{3,4}:

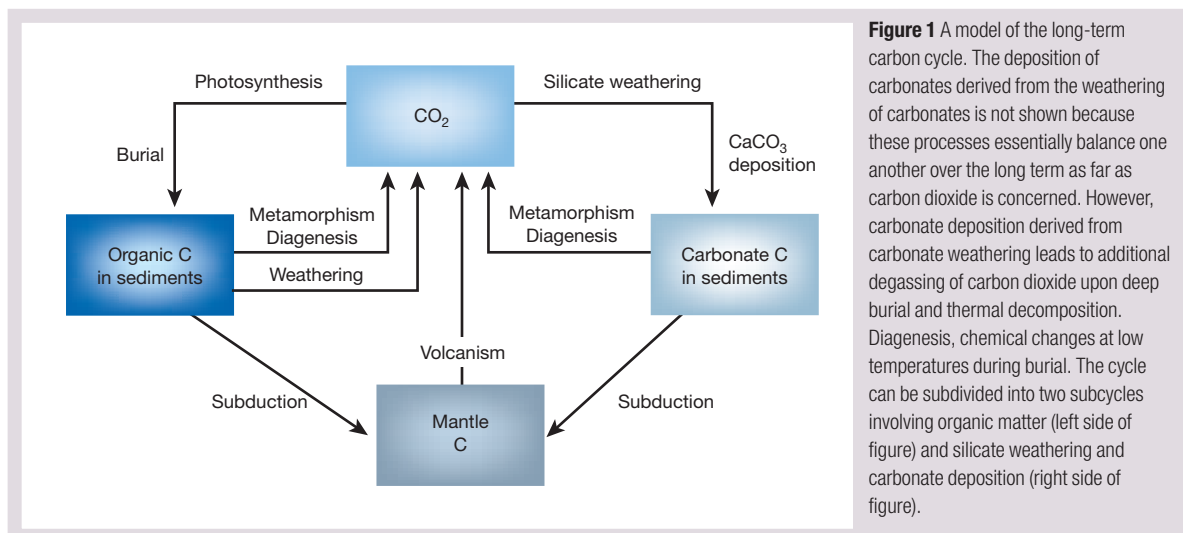


These reactions summarize many intermediate steps. Equation (1), going from left to right, represents the uptake of atmospheric carbon dioxide during the weathering on land of calcium (and magnesium) silicates, with the dissolved weathering products (Ca^{2+} , Mg^{2+} and HCO_3^-) delivered to the ocean and precipitated there as calcium and magnesium carbonates in sediments. Going from right to left, equation (1) represents the deep burial and thermal decom-

position of carbonates with the liberated carbon dioxide released into the atmosphere and oceans.

Equation (2) is of particular interest for this review. Going from left to right, it represents net global photosynthesis (photosynthesis minus respiration), as manifested by the burial of organic matter (CH_2O) in sediments. The buried organic matter is eventually transformed, mostly to kerogen, but some becomes oil, gas and coal. Equation (2), going from right to left, represents the oxidative weathering of organic matter exposed by erosion on the continents (and to a lesser extent the combination of thermal decomposition of organic matter to reduced gases with consequent oxidation of the gases in the atmosphere). The burning of fossil fuels by humans has caused a large increase in the rate of organic matter oxidation compared with that of the natural weathering process. This increase, on the basis of the data we present in this review and that of the IPCC⁵, is an acceleration by a factor of about 100. Thus, humans have greatly perturbed the long-term carbon cycle.

This review shows that numerical modelling, based on the long-term carbon cycle, allows the calculation of global rates of organic matter burial over geological time, and that the resulting rate distribution bears a resemblance to times of formation of oil and gas source rocks and coal. In addition, the complex feedbacks that govern organic burial are shown to be intimately tied to oceanic nutrients, the carbon dioxide and oxygen content of the atmosphere, and climate.



Organic matter burial and source rocks

There are two principal methods for estimating the global rate of burial of organic matter over a geological timescale. The first involves determining the mean concentration of organic carbon in sedimentary rocks and correcting for the loss of the volume of rocks with time owing to erosion, metamorphism (transformation by natural agencies such as heat and pressure) and subduction (the sideways and downwards movement of plates of the Earth's oceanic crust)⁶. The second, reported here, is based on the use of the carbon isotopic composition of seawater. During photosynthesis ^{12}C is taken up preferentially instead of ^{13}C , such that the organic matter derived from organisms is about 20‰ lower in ^{13}C than the carbon in atmospheric carbon dioxide or that in dissolved inorganic carbon (DIC) of sea water. A compilation of $\delta^{13}\text{C}$ of carbonate fossils versus time has been performed⁷ (where $\delta^{13}\text{C}(\text{‰}) = [(^{13}\text{C}/^{12}\text{C}_{\text{sample}})/(^{13}\text{C}/^{12}\text{C}_{\text{standard}}) - 1] \times 1,000$). Changes in $\delta^{13}\text{C}$ of carbonates with time are assumed to record the average $\delta^{13}\text{C}$ value for DIC in sea water and also for atmospheric carbon dioxide (which is about 7‰ lower in ^{13}C than DIC). The rate of preferential removal from the ocean and atmosphere of isotopically light organic carbon, relative to total carbon, is recorded by the value

of $\delta^{13}\text{C}$. For example, an increase in $\delta^{13}\text{C}$ for sea water normally represents increased removal of light carbon and, thus, greater burial of organic matter in sediments. By performing mass-balance calculations for both total carbon and ^{13}C for input and output of carbon to the oceans plus the atmosphere, it is possible, by using the $\delta^{13}\text{C}$ isotopic data, to calculate the global rate of burial, over time, of organic carbon in sediments, both of marine and non-marine origin^{2,8–11}.

A plot of global organic burial rate during the Phanerozoic eon is shown in Fig. 2, and the pattern is a crude guide to the occurrence of fossil-fuel source rocks, as will be seen later. The burial curve is calculated using variable isotope fractionation during photosynthesis^{11,12}, new data on $\delta^{13}\text{C}$ for the late Permian period¹³ and data on the rapid recycling of carbon in younger rocks^{6,11}. The estimated error is $\pm 50\%$ of the rate plotted. The most prominent feature is a broad maximum centred around 300 Myr ago (Permo-Carboniferous period). This maximum is a direct consequence of the very high $\delta^{13}\text{C}$ of the ocean and atmosphere at this time⁷. Prominent minima occur before ~470 Myr ago (early Paleozoic era) and around 220–180 Myr ago (Triassic–Early Jurassic period), whereas secondary maxima are at 440–410 Myr ago (Silurian period) and 90–120 Myr ago (Middle Cretaceous period).

The times of maximum model-calculated organic carbon burial coincide with independent estimates of the times of major oil source rock deposition. The relative areas of the six principal source-rock periods¹⁴ expressed as a percentage of the total for all areas coincide in time with the maxima in the calculated carbon burial curve (Fig. 2). Also, the minima in the carbon-burial curve coincide with periods of very low occurrence of source rocks for oil and coal^{14,15}. This agreement suggests that a major factor in oil generation is increased global deposition of organic matter.

Klemme and Ulmishek¹⁴ have divided their six source-rock periods into two categories, source rocks dominated by type I and type II kerogen (sedimentary organic matter that is high in hydrogen and low in oxygen derived mainly from aquatic planktonic organisms), and those dominated by type III kerogen (lower in hydrogen, higher in oxygen and derived mainly from plants), with coal included with type III as a source rock (mainly for gas). The ratio of type III kerogen to total kerogen for each period is plotted in Fig. 3. The broad maximum centred about 300 Myr ago illustrates the importance of coal deposition at that time. The rise of land plants between 380 and 350 Myr ago during the Devonian period created a new source of relatively non-biodegradable organic matter, lignin, which, when buried in sediments, gave rise to increased global burial of organic carbon in the Carboniferous and Permian (350–250 Myr ago) (Fig. 2). Burial of lignin and other plant products gave rise to the great Permian and Carboniferous coal basins and, to a lesser extent, type III kerogen for oil and gas formation. By far the most abundant coal reserves are from the Permo-Carboniferous period¹⁵, even though these older rocks have been subjected to erosion for longer times than younger deposits. Other periods of high type III kerogen formation (Fig. 3) are the Middle Cretaceous (90–120 Myr ago) and mid-Tertiary (30–50 Myr ago). These are also periods of abundant coal deposition¹⁵. It is notable that the time when both global organic burial and source rock deposition were at their lowest (centred around 200 Myr ago in Fig. 2) is also the time when very little coal was deposited¹⁵.

The temporal distribution of the different kerogen types correlates somewhat with the loci of organic matter deposition over time. The latter can be estimated by means of isotope mass-balance modelling for both carbon and sulphur⁸. This is because of the different sulphur content, which is mainly found in pyrite (FeS_2), of organic-rich sediments deposited in different sedimentary environments. The global rate of burial of sedimentary pyrite can be calculated, in a manner analogous to that used for the burial of organic matter. The sulphur isotopic composition of sea water, as recorded in deposits of gypsum and anhydrite and traces of sulphate in carbonates¹⁶, generally reflect changes in the rate of preferential removal of ^{32}S from sea water as pyrite. Pyrite is lower in $^{34}\text{S}/^{32}\text{S}$ relative to seawater sulphate,

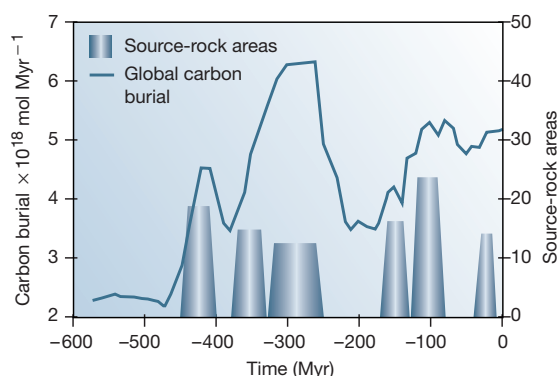


Figure 2 Plot of global organic carbon burial during the Phanerozoic eon compared with the times of deposition of major oil and gas source rocks. (Carbon burial rate modified from ref. 11; source-rock areas from ref. 14.) Units for source-rock area are in percentages of the total of the six areas. Source-rock abundances at other times are very minor¹⁴. The very high carbon burial values centred around 300 Myr ago are due predominantly to terrestrial carbon burial and coal formation⁸.

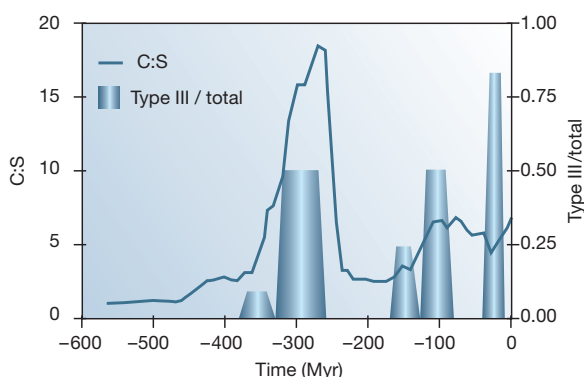


Figure 3 Plots of organic carbon/pyrite sulphur (C:S) buried in sediments versus time, compared with the fraction of oil and gas source rocks that are of type III (including coal). (C:S values from ref. 11; source rock data from ref. 14).

because the reduction of sulphate to H_2S , from which pyrite forms, involves a large isotopic fractionation¹⁶. Sulphur isotopic data for the ocean, combined with mass-balance calculations for total sulphur and ^{34}S , enable the calculation of rates of pyrite burial over time.

Ratios of the burial rates of organic carbon and pyrite sulphur (C:S) versus time are also shown in Fig. 3. The changes with time of C:S agree with what is known about pyrite formation in modern sediments⁸. Predominant burial in organic-rich marine sediments, for instance in basins with anoxic bottom water such as the Black Sea, involves abundant sulphate reduction and pyrite formation, leading to low C:S ratios. Burial in fresh-water environments, such as coal swamps, results in very high C:S ratios. This is because of the very low sulphate content of fresh water compared with that of sea water, which minimizes pyrite formation⁸. (Burial in oxygenated marine bottom waters with bioturbation results in intermediate C:S values, but this environment is not as favourable for oil formation¹⁷). Thus, for oil-prone marine source rocks deposited under anoxic bottom-water conditions¹⁷, one would expect low C:S ratios, and for coal deposits, one would expect high C:S ratios. The correlation between type III kerogen, relative to total kerogen, and C:S ratio (Fig. 3) agrees with this prediction for several source-rock periods (Fig. 3).

The very low values of C:S for the early Paleozoic (before 370 Myr ago) suggest greater deposition in unusually organic-rich and pyrite-rich marine rocks. This is in agreement with the widespread occurrence of the graptolite black shale facies of this era, which has been interpreted to have been deposited in abundant anoxic regions of the ocean¹⁸. (The very low C:S values also reflect the relative lack of deposition in fresh water, owing to the absence of large land plants at that time.) The argument that there is little difference in organic preservation between oxygenated and non-oxygenated bottom waters¹⁹ is controversial and does not agree with other studies of modern sediments or with geological observations^{17,20,21}.

The presence of only a moderately high proportion of type III kerogen in Permo-Carboniferous source rocks, compared with the very high concurrent values of C:S, may be because the compilation of Klemme and Ulmishek¹⁴ was oriented towards oil and gas formation, not coal abundance. Also, the poor correlation between the ratio of type III to total kerogen and C:S for the mid-Tertiary (40–20 Myr ago) is explained by the deposition of much terrestrially derived (type III) organic matter in marine estuarine and deltaic environments¹⁴, where abundant pyrite formation is supported by sulphate-rich sea water.

Organic burial, nutrients and atmospheric composition

The burial of organic matter, as it affects source-rock formation, is intimately tied to the chemical composition of the oceans and atmosphere. For the atmosphere, as shown by equation (2), the burial of organic matter involves the removal of carbon dioxide and the addition of oxygen. In this way, organic burial, on a geological timescale, is a major controller of atmospheric composition. Atmospheric composition, in turn, exerts a major control on organic matter burial and acts in a feedback capacity. A major factor linking organic burial in the marine environment and atmospheric composition is the level of essential nutrients in sea water, principally phosphorus and nitrogen²². This is because biological production is normally limited by nutrients. Changes in atmospheric carbon dioxide and oxygen can affect the availability of nutrients, which in turn affect organic production and burial. This leads to a variety of biogeochemical feedbacks that control both atmospheric composition and organic burial. Because of this complex web of feedbacks, a simple explanation of the various changes over time in rates of organic burial and source-rock formation (Figs 2 and 3) is difficult, with the exception of the effect of the rise of large land plants on terrestrial carbon burial.

The interaction of nutrients and atmospheric carbon dioxide and oxygen with organic burial can be represented by systems-analysis diagrams. An example is shown in Fig. 4. (A complete coverage of all suggested linkages is beyond the scope of this review — for more

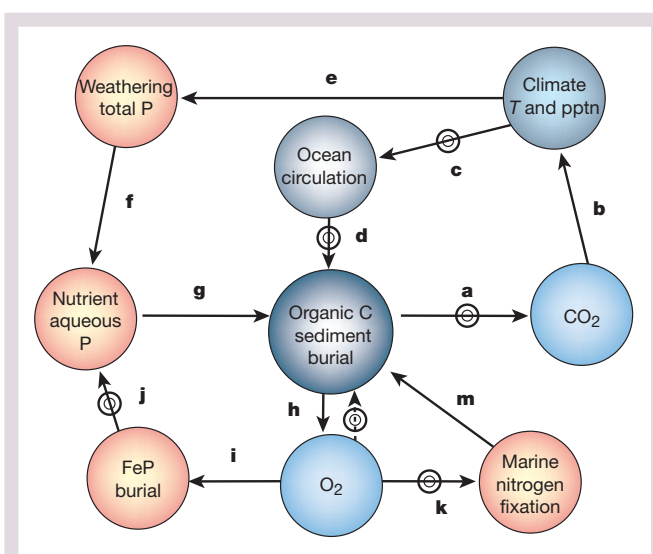


Figure 4 Systems-analysis diagram showing some of the feedback relationships between marine organic carbon burial, nutrients, climate, atmospheric carbon dioxide and oxygen, and ocean circulation. The upward dashed arrow at **h** represents disagreement concerning the effect of oxygen on organic burial. (See text for detailed explanation). Solid arrows represent positive responses and arrows with bullseyes represent negative responses. A loop with an even number of arrows with bullseyes, or with no bullseyes, represents a positive feedback loop. A loop with an odd number of bullseyes represents negative feedback. FeP, phosphorus associated with ferric oxide.

details consult refs 23–25.) Plain arrows between entities represent positive responses. This means, for example, if organic burial increases, oxygen production and, therefore oxygen level, increases. Arrows with attached bullseyes refer to negative responses. For example, if organic burial increases, carbon dioxide (the ultimate source of the carbon) decreases. By following the diagram, one can deduce whether various loops lead to positive or negative feedback. A loop with an even number of arrows with bullseyes or no bullseyes represents positive feedback. A loop with an odd number of bullseyes represents negative feedback. Because atmospheric oxygen and carbon dioxide have not varied considerably over geological timescales, interest has focused on negative feedback pathways, which provide stabilization.

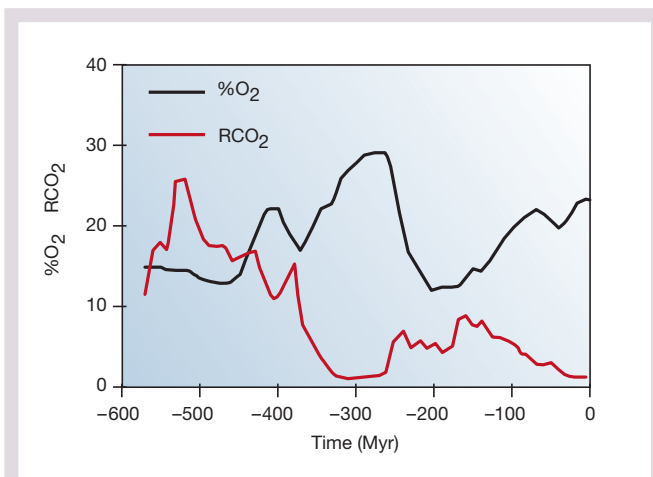


Figure 5 Plots of RCO_2 (the ratio of the mass of carbon dioxide in the atmosphere in the past to that for the pre-industrial present) and $\%\text{O}_2$ during the Phanerozoic eon. (Values of RCO_2 from the GEOCARB III model³³; values of $\%\text{O}_2$ from ref. 11 using the ^{13}C data of refs 12 and 13). Estimated errors are $\pm 50\%$ for RCO_2 and $\pm 7\%$ for O_2 .

It is instructive to follow several of the loops shown in Fig. 4. Consider the loop h–i–j–g. Increased organic burial raises atmospheric oxygen, which results in a more oxygenated ocean and the greater burial of phosphorus associated with ferric oxides (FeP). Burial of FeP robs the ocean of available dissolved phosphorus for photosynthesis, resulting in a lower mean phosphorus concentration, and this results in lowered biological productivity and less organic burial. This negative feedback loop is thought to be a major stabilizing control on atmospheric oxygen over a geological timescale^{26,27}. An alternative suggestion is that nitrogen fixation is the primary factor limiting organic production²⁸. In this case, loop h–k–m provides the negative feedback; greater organic burial produces higher oxygen, and this inhibits biological nitrogen fixation, which means lower nitrogen levels in sea water and lower organic production and burial.

Atmospheric carbon dioxide is also involved in negative feedback with regard to organic carbon burial. Consider the loop a–b–c–d. Greater burial leads to a lowering of atmospheric carbon dioxide, which, because of the atmospheric greenhouse effect, leads to global cooling, especially at high latitudes²⁹. A cooler Earth with a greater latitudinal temperature gradient could drive greater oceanic circulation and a greater supply of oxygen to deep waters. This inhibits the formation of anoxic bottom waters and thereby reduces organic burial¹⁷. This loop has three arrows with bullseyes, and this indicates overall negative feedback. (An additional scenario is that greater oceanic circulation could lead to a greater supply of nutrients to surface waters with greater organic production, greater burial and positive feedback). Finally, there is the loop a–b–e–f–g. Higher organic burial causes lower carbon dioxide and a cooler Earth, which reduces the rate of weathering on land of rocks containing phosphorus. This means less delivery of phosphorus to the oceans, lower marine-dissolved phosphorus, and lower biological production and organic burial^{30,31}. Other possible negative-feedback loops have been suggested, such as greater organic burial, elevated oxygen, reduced plant-assisted weathering of phosphate minerals, less phosphorus transport to the sea and lower organic carbon burial²³.

The arrow going from oxygen to organic matter is dashed. This is because there is controversy over whether the oxygen content of the atmosphere and oceans has a direct influence on global organic matter burial. Studies of the modern ocean^{20,21} suggest that organic matter preservation in sediments is considerably reduced by the presence of oxygen in bottom waters, but a simple inverse correlation between measurable oxygen level and organic matter burial has not been found³². Also, high concentrations of organic matter can be found in oxygenated bottom waters under zones of high plankton productivity¹⁹.

Using a large variety of feedback mechanisms (many of which are not illustrated in Fig. 4), concentrations of carbon dioxide and oxygen in the atmosphere over a geological timescale have been calculated by computer modelling^{6,11,31,33–36}. A review of these models and their results can be found in ref. 25. An example based on modelling, by myself, of a variety of processes^{11,33}, including calcium and magnesium silicate and carbonate weathering and burial, organic matter weathering and burial, pyrite weathering and burial, and global CO₂ degassing, is shown in Fig. 5. (From sensitivity analysis, the errors in carbon dioxide are about $\pm 50\%$ and for oxygen about $\pm 7\%$ oxygen.) A variety of independent estimates agree with the curve for carbon dioxide^{37,38}, and biological data agree with elevated Permo-Carboniferous oxygen levels²⁴. The most prominent feature in these plots are the high oxygen levels and low carbon dioxide levels during the Permo-Carboniferous. These are due primarily to the rise of large vascular land plants and their effect on accelerated silicate weathering (for carbon dioxide) and increased organic matter burial (for both carbon dioxide and oxygen).

Modelling of the long-term carbon cycle can lead to new insights regarding the formation of source beds for oil, coal and gas and

the relationship between global organic matter burial and the composition of the atmosphere. Further work is needed to better identify the principal feedback relations and to better quantify the various fluxes within the carbon cycle. □

doi:10.1038/nature02131

- Garrels, R. M., Lerman, A. & Mackenzie, F. T. Controls of atmospheric O₂ and CO₂ — past, present, and future. *Am. Sci.* **64**, 306–315 (1976).
- Holland, H. D. *The Chemistry of the Atmosphere and Oceans* (Wiley, New York, 1978).
- Ebelmen J. J. Sur les produits de la décomposition des espèces minérales de la famille des silicates. *Annu. Rev. Mines* **12**, 627–654 (1845).
- Urey, H. C. *The Planets: their Origin and Development* (Yale Univ., New Haven, 1952).
- IPCC (Intergovernmental Panel on Climate Change) *Climate Change 2001: Synthesis Report* (IPCC, Geneva, 2001).
- Berner, R. A. & Canfield, D. E. A model for atmospheric oxygen over Phanerozoic time. *Am. J. Sci.* **289**, 333–361 (1989).
- Veizer, J. et al. ⁸⁷Sr/⁸⁶Sr, δ^{13} C and δ^{18} O evolution of Phanerozoic seawater. *Chem. Geol.* **161**, 59–88 (1999).
- Berner, R. A. & Raiswell, R. Burial of organic carbon and pyrite sulfur in sediments over Phanerozoic time: a new theory. *Geochim. Cosmochim. Acta* **47**, 855–862 (1983).
- Garrels R. M. & Lerman, A. Coupling of the sedimentary sulfur and carbon cycles — an improved model. *Am. J. Sci.* **284**, 989–1007 (1984).
- Walker J. C. G. Global geochemical cycles of atmospheric oxygen. *Mar. Geol.* **70**, 159–174 (1986).
- Berner, R. A. Modeling atmospheric O₂ over Phanerozoic time. *Geochim. Cosmochim. Acta* **65**, 685–694 (2001).
- Hayes, J. M., Strauss, H. & Kaufman, A. J. The abundance of ¹³C in marine organic matter and isotope fractionation in the global biogeochemical cycle of carbon during the past 800 Ma. *Chem. Geol.* **161**, 103–125 (1999).
- Korte, C., Kozur, H. W., Joachimski, M. M. & Veizer, J. Strontium, oxygen and carbon isotope records of Permian seawater. *Eur. Geophys. Soc. Geophys. Res. Abstr.* **5**, 13061 (2003).
- Klemme, H. D. & Ulmishek, G. F. Effective petroleum source rocks of the world: stratigraphic distribution and controlling depositional factors. *Bull. Am. Ass. Petrol. Geol.* **75**, 1809–1851 (1991).
- Bestougeff, M. A. Summary of world coal resources and reserves. *26th Int. Geol. Cong. Paris Colloq.* **35**, 353–366 (1980).
- Strauss, H. Geological evolution from isotope proxy signals — sulfur. *Chem. Geol.* **161**, 89–101 (1999).
- Demailson, G. J. & Moore, G. T. Anoxic environments and oil source bed genesis. *Bull. Am. Ass. Petrol. Geol.* **64**, 1179–1209 (1980).
- Berry, W. B. N. & Wilde, P. Progressive ventilation of the oceans — an explanation for the distribution of the lower Paleozoic black shales. *Am. J. Sci.* **278**, 257–275 (1978).
- Pederson, T. F. & Calvert, S. E. Anoxia vs productivity: what controls the formation of organic carbon-rich sediments and sedimentary rocks. *Bull. Am. Ass. Petrol. Geol.* **74**, 454–466 (1990).
- Canfield, D. E. Factors influencing organic carbon preservation in marine sediments. *Chem. Geol.* **114**, 315–329 (1994).
- Hedges J. I. et al. Sedimentary organic matter preservation: A test for selective degradation under oxic conditions. *Am. J. Sci.* **299**, 529–555 (1999).
- Falkowski, P. G. & Raven, J. *Aquatic Photosynthesis* (Blackwell, Oxford, 1997).
- Lenton T. M. & Watson, A. J. Redfield revisited 2: What regulates the oxygen content of the atmosphere? *Glob. Biogeochem. Cycles* **14**, 249–268 (2000).
- Berner, R. A., Beerling, D. J., Dudley, R., Robinson, J. M. & Wildman, R. A. Phanerozoic atmospheric oxygen. *Annu. Rev. Earth Planet. Sci.* **31**, 105–134 (2003).
- Berner, R. A. *The Phanerozoic Carbon Cycle* (Oxford Univ. Press, Oxford, in the press).
- Van Cappellen, P. & Ingall, E. D. Redox stabilization of the atmosphere and oceans by phosphorus-limited marine productivity. *Science* **271**, 493–496 (1996).
- Colman, A. S. & Holland, H. D. *Marine Authigenesis: From Microbial to Global* (eds Glenn, C., Lucas, J. & Prevot-Lucas, L.) 53–75 (Soc. Econ. Paleontologists & Mineralogists, 2000).
- Falkowski, P. G. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO₂ in the ocean. *Nature* **387**, 272–275 (1997).
- Crowley, T. J. & North, G. R. *Paleoclimatology* (Oxford Univ. Press, Oxford, 1991).
- Weissert, H. C. Isotope stratigraphy, a monitor of paleoenvironmental change: a case study from the Early Cretaceous. *Surv. Geophys.* **10**, 1–61 (1989).
- Wallmann, K. Controls on the Cretaceous and Cenozoic evolution of seawater composition, atmospheric CO₂ and climate. *Geochim. Cosmochim. Acta* **65**, 3005–3025 (2001).
- Betts, J. N. & Holland, H. D. The oxygen content of ocean bottom waters, the burial efficiency of organic-carbon and the regulation of atmospheric oxygen. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **97**, 5–18 (1991).
- Berner, R. A. & Kothavala, Z. GEOCARB III: A revised model of atmospheric CO₂ over Phanerozoic time. *Am. J. Sci.* **301**, 182–204 (2001).
- Hansen, K. W. & Wallmann, K. Cretaceous and Cenozoic evolution of seawater composition, atmospheric O₂ and CO₂. *Am. J. Sci.* **303**, 94–148 (2003).
- Tajika, E. Climate change during the last 150 million years: reconstruction from a carbon cycle model. *Earth Planet. Sci. Lett.* **160**, 695–707 (1998).
- Bergman, N., Lenton, T. & Watson, A. Coupled Phanerozoic predictions of atmospheric oxygen and carbon dioxide. *Eur. Geophys. Soc. Geophys. Res. Abstr.* **5**, 11208 (2003).
- Royer, D. L., Berner, R. A. & Beerling, D. J. Phanerozoic atmospheric CO₂ change: evaluating geochemical and paleobiological approaches. *Earth Sci. Rev.* **54**, 349–392 (2001).
- Crowley, T. J. & Berner, R. A. CO₂ and climate change. *Science* **292**, 870–872 (2001).

Acknowledgements This research was supported by grants from the US Department of Energy and the US National Science Foundation.

Competing interests statement The author declares that he has no competing financial interests.

Organic–inorganic interactions in petroleum-producing sedimentary basins

Jeffrey S. Seewald

Department of Marine Chemistry and Geochemistry, MS #4, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA (e-mail: jseewald@whoi.edu)

Petroleum deposits form as a consequence of the increased temperatures that accompany progressive burial of organic matter deep within sedimentary basins. Recent advances in petroleum geochemistry suggest that inorganic sedimentary components participate in organic transformations associated with this process. Water is particularly important because it facilitates reaction mechanisms not available in dry environments, and may contribute hydrogen and oxygen for the formation of hydrocarbons and oxygenated alteration products. These findings suggest that petroleum generation and stability is influenced by subsurface chemical environments, and is a simple function of time, temperature and the composition of sedimentary organic matter.

Petroleum consists of liquid (oil) and volatile (natural gas) organic alteration products, and is generated during the conversion of metastable macromolecular kerogen to thermodynamically favoured lower molecular mass compounds.

The chemical reactions responsible for this transformation occur in response to the removal of kinetic barriers as temperatures increase with progressive burial in sedimentary basins. Formation of an economic petroleum deposit requires, in addition to suitable source rocks containing sufficient organic matter, a sequence of geological events that leads to the expulsion, migration and trapping of generated hydrocarbons. Despite great success in finding and producing petroleum reserves during the past century, there is still extensive debate regarding the chemical and physical processes that result in the generation and accumulation of oil and natural gas.

Traditional models for the formation of oil and natural gas typically involve a catagenic process dominated by thermal cracking reactions that release low-molecular-mass hydrocarbon fragments^{1,2}. These reactions are viewed as unidirectional, kinetically controlled processes that are influenced solely by time, temperature and the composition and structural characteristics of the source kerogen. In recent years, however, geochemists have increasingly recognized that subsurface chemical environments play an active part in the formation and compositional evolution of petroleum. In particular, inorganic compounds such as water and minerals may participate as reactants or catalysts during organic matter maturation. Moreover, organic–inorganic interactions in sedimentary basins have direct implications for petroleum migration and trapping, because many organic alteration products participate in processes that create or destroy sediment porosity and permeability.

Here, I focus on the recent advances in our understanding of organic–inorganic interactions that have the greatest potential to influence the generation, chemical evolution and accumulation of petroleum. A variety of field, laboratory and theoretical approaches have been adopted in studies that identify a broad spectrum of interactions involving water, minerals and hydrocarbons during the formation of a petroleum deposit. Many of these studies are controversial, and few have gained general acceptance. Nonetheless, they represent new paradigms that may substantially enhance our ability to predict the occurrence of petroleum.

Petroleum generation and expulsion

In most sedimentary basins, the onset of petroleum generation begins at approximately 50 °C, at depths greater than 1 km (refs 1, 2). The overall process involves the initial cleavage of weak non-covalent bonds in kerogen to form a macromolecular bitumen phase, followed by cleavage of covalent bonds in bitumen to form liquid oil and natural gas^{3,4} (Fig. 1). In some cases, oil and gas may also form directly from kerogen without a bitumen intermediary^{3,5,6}. According to conventional wisdom, continued heating of oil results in its conversion to natural gas, although the extent to which this actually takes place in nature has been the focus of much debate. The broad and continuous size range of the organic molecules that compose petroleum reflect the diversity of structural moieties in macromolecular kerogen and bitumen precursors.

The reaction mechanisms responsible for the generation of petroleum play a significant role in regulating petroleum's composition. Carbonium-ion mechanisms (those involving positively charged intermediaries), for example, would be expected to produce a predominance of branched products, whereas free-radical mechanisms (those involving neutrally charged intermediaries) favour straight-chain products⁷. The predominance of straight-chain alkanes in petroleum supports the free-radical mechanism, although a carbonium-ion mechanism may be important for the production of branched low-molecular-mass compounds in immature sediments that have not yet undergone extensive petroleum generation⁸. A free-radical mechanism is further supported by observations from field and laboratory studies that sulphur-rich kerogens generate oil at lower thermal stress than sulphur-poor kerogens^{9–12}. This trend has been attributed to abundant sulphur radicals released from kerogen that initiate carbon–carbon bond cleavage¹³.

Mineral catalysis has been suggested as an important factor for petroleum generation on the basis of the catalytic activity of acidic clays during refinery processing^{14–16}. The extent to which clays participate in acid-catalysed transformations in petroleum generation, however, is dependent on their catalytic activity under the physical and chemical conditions that are likely to exist in geological environments. Specifically, the catalytic activity of clays is suppressed in the presence of water^{15,17}, an abundant component of sedimentary rocks. Moreover, reaction mechanisms associated with the catalytic activity of clays involve carbonium-ion

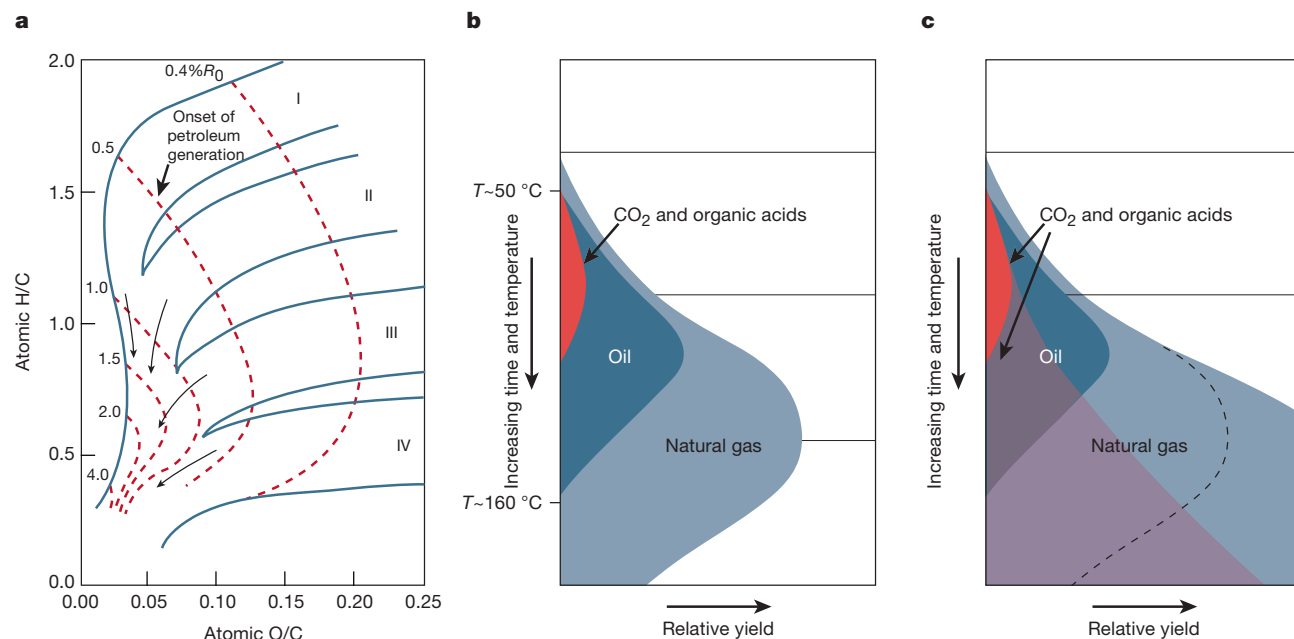


Figure 1 Chemical evolution of kerogen and petroleum during thermal maturation in sedimentary basins. **a**, Van Krevelen diagram showing the chemical evolution of immature kerogen of varying composition (type I, II, III and IV) at increasing levels of thermal maturity (based on ref. 2). Levels of thermal maturity are indicated by isochors of vitrinite reflectance (R_0), a widely used geochemical indicator that integrates the effects of time and temperature during thermal maturation of sediments. In general, kerogen composition moves from the upper right regions of the figure to the lower left with increasing maturity. **b**, Traditional model of the amount and timing of organic alteration products generated during progressive burial

in sedimentary basins that assumes oxygen and hydrogen in organic alteration products are derived only from kerogen (modified from ref. 2). The form of this figure is constrained by the maturation trends shown in the Van Krevelen diagram. **c**, Schematic illustration of the amount and timing of organic alteration products generated if water and minerals are allowed to contribute the requisite hydrogen and oxygen for the formation of hydrocarbons and oxygenated compounds such as carbon dioxide and carboxylic acids. Ultimately, production of oxygenated products and methane will cease owing to exhaustion of a reactive carbon source. The depth at which this occurs is unknown.

intermediaries that generate a predominance of branched products. Thus, although clay catalysis can occur in subsurface environments devoid of water, the relatively low concentrations of branched alkanes in petroleum suggests it is not a major process.

Chemical reactions that regulate petroleum generation are often obscured by the complexity of natural systems and cannot be determined solely from field studies. Accordingly, experiments that heat petroleum source rocks under well-constrained temperature, pressure and chemical conditions are routinely used to simulate natural processes and identify key variables during kerogen maturation. Laboratory experiments have demonstrated that kerogen decomposition to form bitumen is not influenced by the presence of water⁴. By contrast, water influences the decomposition of bitumen by retarding secondary cracking of long-chain hydrocarbons to produce gas, and condensation reactions that result in the formation of highly aromatic carbon-rich residues commonly referred to as pyrobitumen or char^{4,17,18}. The net result is an increased yield and increased stability for hydrocarbons generated from source rocks in the presence of water⁴. This effect suggests that molecular hydrogen derived from water dissolved in bitumen quenches transitory free-radical sites that form on the kerogen or bitumen macromolecule^{4,18}.

A critical question associated with the role of water during petroleum generation is to what extent water-derived hydrogen and oxygen contribute to the formation of hydrocarbons and oxygenated alteration products. Isotopically labelled experiments during which source rocks were heated in the presence of liquid water (hydrous pyrolysis) have provided compelling evidence for the incorporation of water-derived hydrogen and oxygen in saturated alkanes and alkylphenols, respectively, during petroleum generation^{19–21}. A common feature of hydrous pyrolysis experiments is the production of large quantities of carbon dioxide^{4,22–25}. In carbonate-free experiments, sources of

carbon dioxide are limited to alteration of the initial kerogen and oxidation of organic carbon compounds. The amount of oxygen in carbon dioxide produced during these experiments exceeds the oxygen content of the initial kerogen, providing direct evidence that water-derived oxygen contributes to the formation of carbon dioxide^{4,25}.

The spatial separation of most petroleum reservoirs from source rocks indicates that large-scale migration of oil and natural gas must take place. The initial expulsion from fine-grained source rocks, referred to as primary migration, releases hydrocarbons into permeable conduits, where secondary migration may result in buoyancy-driven transport to reservoirs. Creation of elevated pressures within the pore spaces of source rocks is generally believed to be the major cause of primary expulsion, although gaseous solution and other processes may contribute^{1,2}. Models suggested for the creation of pressure gradients involve either compaction of sediments during progressive burial²⁶ or volume expansion accompanying petroleum generation²⁷. Hunt *et al.*²⁷ have pointed out that at depths associated with petroleum generation, many source rocks do not undergo compaction, and suggest that the latter process is responsible for primary migration. The ability of source rocks to expel oil in the absence of compaction during hydrous pyrolysis experiments also supports an active role for volume expansion during primary migration⁴.

Extensive chemical fractionation occurs during primary migration of oil. Whereas expelled oil is rich in saturated hydrocarbons, the bitumen that remains in source rocks contains abundant polar compounds such as resins and asphaltene and is depleted in saturated compounds¹. Selective adsorption of polar compounds on mineral surfaces or in kerogen might result in large compositional differences between expelled oil and retained bitumen, but laboratory investigations of such processes have not been successful in reproducing the compositional trends for oil and bitumen observed in nature^{28–31}.

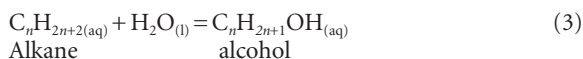
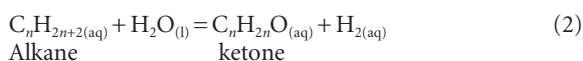
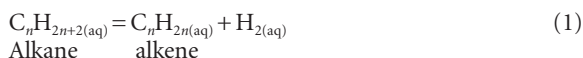
The ability of hydrous pyrolysis experiments to produce an expelled oil that is compositionally similar to natural oils led Lewan⁴ to speculate that the solvent properties of bitumen are altered by the presence of dissolved water, causing formation of an immiscible saturate-rich phase that is subsequently expelled from the rock. Thus, water represents an essential participant in chemical processes associated with both the generation and expulsion of oil from source rocks.

Chemical interactions in migration conduits and reservoirs

Following expulsion, petroleum may encounter substantial quantities of liquid water in migration conduits and reservoirs. The large quantities of aqueous brine co-produced with oil and natural gas indicates a close association between them in reservoir environments. In reservoirs where aqueous pore fluids are displaced by hydrocarbons, an extensive oil–water interface may be maintained by water-coated mineral surfaces^{32,33}. Reactions that can influence the chemical evolution of petroleum in migration conduits and reservoirs involve species dissolved in liquid water or hydrocarbon phases. Aqueous reactions are of particular interest because, in addition to representing an abundant source of hydrogen and oxygen for the formation of hydrocarbons and oxygenated alteration products, water can serve as a solvent and catalyst that facilitates reaction pathways not available in an oil or gas phase. The reactivity of aqueous organic species increases at increased temperatures in response to large changes in the physical and chemical properties of water. In particular, the density and dielectric constant decrease with increasing temperature, whereas the dissociation constant increases³⁴. Increases in the dissociation constant produce higher H^+ and OH^- activities, which can enhance both acid- and base-catalysed reactions³⁴.

Some may question the significance of reactions occurring in water on the grounds that the low solubility of long-chain hydrocarbons limits the ability of aqueous processes to affect the composition of a coexisting hydrocarbon phase. Microbial degradation of oil, however, is a common process occurring at the oil–water interface and causes the disappearance of entire compound classes³⁵. Similarly, thermochemical sulphate reduction, which involves the oxidation of oil by sulphate-bearing evaporite minerals, is an aqueous process that can extensively alter the composition of oil. In extreme cases, water washing is an aqueous process whereby large volumes of water selectively remove the more water-soluble components of petroleum³⁵. These observations indicate that the rate of aqueous organic reactions and rate of partitioning of organic species between oil and water are sufficiently rapid to influence the composition of a coexisting hydrocarbon phase.

Numerous laboratory experiments have demonstrated that hydrocarbons dissolved in liquid water are highly reactive at increased temperatures and pressures^{34,36–38}. In experiments with iron-bearing mineral assemblages that buffer the activity of aqueous hydrogen, the relative abundances of alkanes and their corresponding alkenes, ketones and alcohols rapidly attain a state of reversible metastable thermodynamic equilibrium according to the following redox-dependent reactions^{36,38}:



The term metastable thermodynamic equilibrium refers to a chemical state in which kinetic barriers prevent the system from attaining total thermodynamic equilibrium, while, at the same time, individual metastable species may reach a state of partial equilibrium

if suitable reaction paths exist. Thermodynamic equilibration according to reactions (1) to (3) does not imply that large amounts of alkenes, ketones and alcohols exist in natural systems. Equilibrium constants for these reactions indicate that in the highly reducing conditions of petroleum reservoirs, equilibrium concentrations of alkanes are several orders of magnitude greater than co-existing alkenes, ketones and alcohols, which is consistent with the levels observed³⁸. Low concentrations notwithstanding, the observation that the relative abundance of aqueous alkenes, ketones and alcohols is regulated by the activity of aqueous molecular hydrogen, which in turn is buffered by the iron-bearing mineral assemblages, is highly significant. It demonstrates that the reversible equilibrium state involves water and iron-bearing minerals. In other words, the exchange of hydrogen and oxygen between aqueous alkanes, alkenes, alcohols, ketones, water and minerals is a rapid process with minimal kinetic barriers.

Direct reaction of aqueous organic species with water during redox-buffered laboratory experiments is similar to hydrolytic disproportionation reactions postulated to control the relative abundance of hydrocarbons in petroleum at the oil–water interface³⁹. Hydrolytic disproportionation involves the reaction of hydrocarbons with water to produce shorter chain carbon compounds that incorporate water-derived hydrogen and oxygen. This type of process can be represented by the reaction:



which demonstrates the generation of alteration products that contain carbon in nominal oxidation states that are more reduced and oxidized than the initial hydrocarbon.

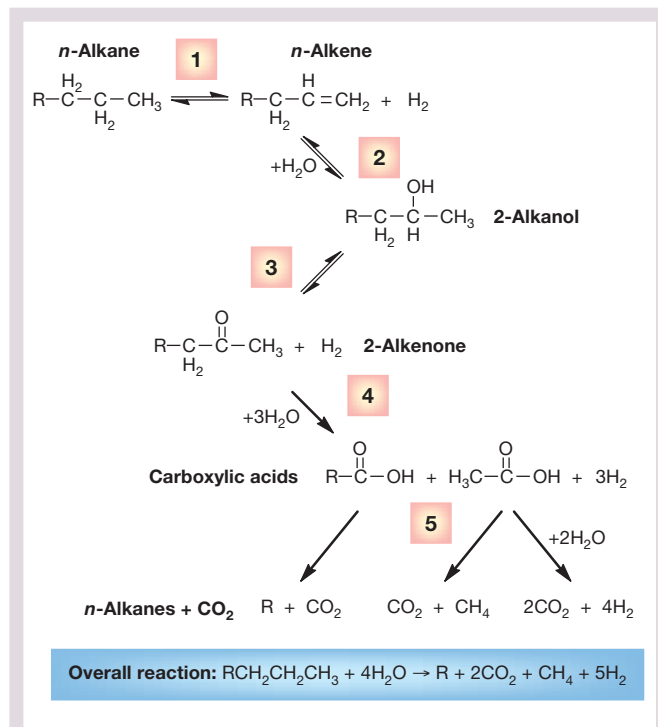


Figure 2 Reaction pathways responsible for the stepwise oxidation of aqueous *n*-alkanes at elevated temperatures and pressures. Mineral oxidants in subsurface environments may consume molecular hydrogen generated by these reactions, allowing the overall reaction to proceed continuously. Saturated hydrocarbons produced in step (5) may re-enter the sequence at the top and undergo subsequent oxidation. The net effect is the conversion of long-chain alkanes in oil to short-chain hydrocarbons in natural gas. The overall reaction indicated at the bottom has been written assuming that decarboxylation is responsible for the decomposition of acetic acid in the final step of the sequence.

Helgeson *et al.*³⁹ concluded from a thermodynamic evaluation of oilfield brine and petroleum compositions that hydrocarbons, aqueous carboxylic acids and carbon dioxide produced by hydrolytic disproportionation may attain a state of reversible metastable thermodynamic equilibrium that includes sedimentary minerals such as calcite, according to the following reactions:

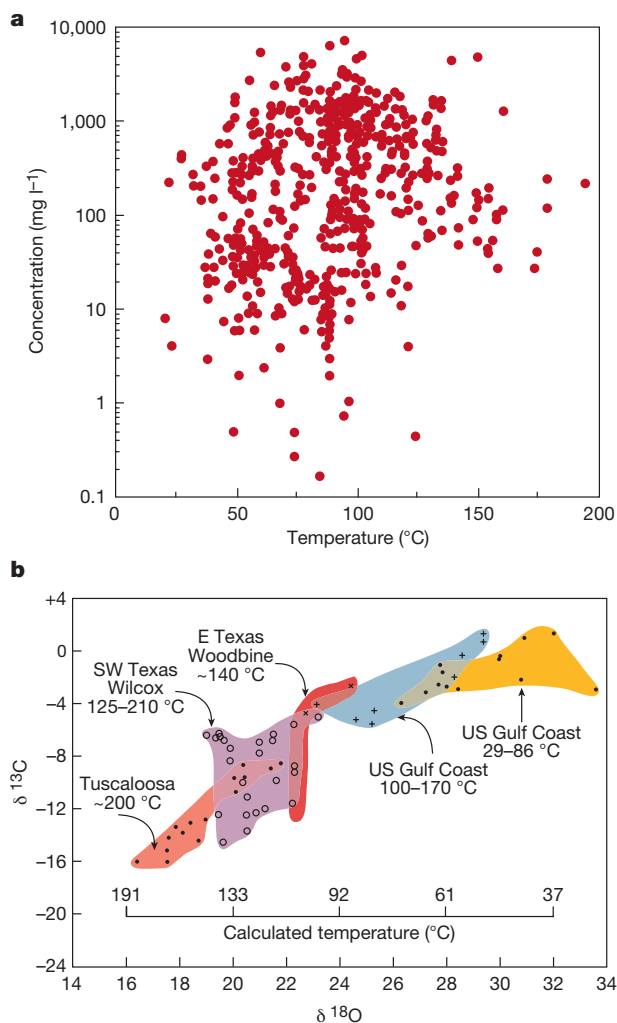
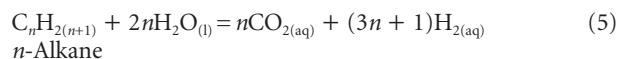
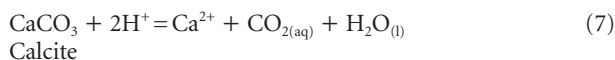


Figure 3 Evidence for the production of oxygenated organic alteration products at levels deep within sedimentary basins. **a**, Concentrations of organic acids in oilfield brines plotted as a function of subsurface temperature (modified from ref. 82). The existence of acids at temperatures in excess of peak petroleum generation (see Fig. 1) is consistent with generation through late-stage reactions involving *n*-alkanes and water. **b**, Carbon and oxygen isotope composition of carbonate cements in US Gulf Coast sedimentary rocks (modified from ref. 58). The trend of decreasing $\delta^{13}C$ with decreasing $\delta^{18}O$ indicates an increased contribution of carbon dioxide containing organically derived (isotopically depleted) carbon at temperatures approaching 200 °C. The temperature scale in **b** was calculated assuming equilibrium ^{18}O fractionation (ref. 83) during formation of carbonate cements from a typical of Gulf Coast formation water characterized by an ^{18}O content of 6‰ (ref. 83). The calculated temperatures are generally consistent with measured downhole temperatures indicated next to each location.



If correct, equilibrium models of this type indicate that reactions between water and organic compounds regulate the composition of petroleum in subsurface environments. Moreover, the relative abundances of species controlled by a state of thermodynamic equilibrium are independent of time, and respond systematically to temperature, pressure and the chemical composition of a given chemical system.

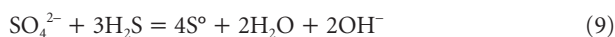
Oxidative processes

Because most organic reactions involve a change in the oxidation state of carbon, redox-reactive sedimentary minerals can influence the stability of petroleum. Thermochemical sulphate reduction (TSR) is a well-documented process that introduces contaminants such as carbon dioxide and hydrogen sulphide to oil and natural gas^{40–43}. TSR involves the reduction of aqueous sulphate derived from the dissolution of anhydrite ($CaSO_4$) in evaporite-bearing sediments during petroleum oxidation to produce hydrogen sulphide, sulphur, calcite and carbon dioxide. Owing to the close association of evaporites with carbonate sequences, TSR is commonly observed in carbonate reservoirs. The overall process can be represented by the general reaction:



The estimated minimum temperatures required for TSR are between 120 °C to 140 °C, although they may be as low as 100 °C (ref. 44). The range of temperatures probably reflects the diverse reaction kinetics associated with the oxidation of compositionally variable petroleum deposits. Large compositional changes are observed in oils affected by TSR^{41,45–47}. In particular, saturated straight-chain hydrocarbons represent a highly reactive component of petroleum that is oxidized during TSR. By contrast, methane is relatively unreactive. Oxidation of petroleum also results in the generation of substantial aromatic and sulphur-bearing compounds.

Laboratory experiments have shown that oxidation of hydrocarbons by sulphate involves the following reactions:



where S^0 represents sulphur in intermediate oxidation states and CH_2 represents reactive hydrocarbons^{40,48–50}. It should be recognized that reactions (9) and (10) are simplifications of complex multi-step chemical processes. Additional experiments examining the oxidation of aqueous *n*-alkanes at increased temperatures and pressures have revealed that oxidation occurs readily in the presence of ferric-iron-bearing mineral assemblages as well as anhydrite³⁸. The oxidation process consists of a series of sequential reactions involving alkene, alcohol, ketone and carboxylic acid intermediaries, and ultimately produces saturated hydrocarbons of shorter chain length and carbon dioxide (Fig. 2). As is the case for TSR, oxidation of *n*-alkanes by iron-bearing minerals is catalysed by intermediate oxidation state sulphur compounds³⁸. In many ways, the overall reaction shown in Fig. 2 is analogous to the hydrolytic disproportionation reactions described above, in that *n*-alkanes react with water to produce shorter chain saturated hydrocarbons and carbon dioxide. The mineral oxidants maintain a strong drive for the reaction to proceed from left to right by the consumption of molecular hydrogen that would otherwise accumulate in solution. During hydrolytic disproportionation, molecular hydrogen is consumed by reduction of organic compounds. The rapid kinetics for reactions that consume alkene, alcohol and ketone intermediaries result in very low concentrations of these species in solution during *n*-alkane oxidation. By contrast, sluggish kinetics for the destruction of carboxylic acids by

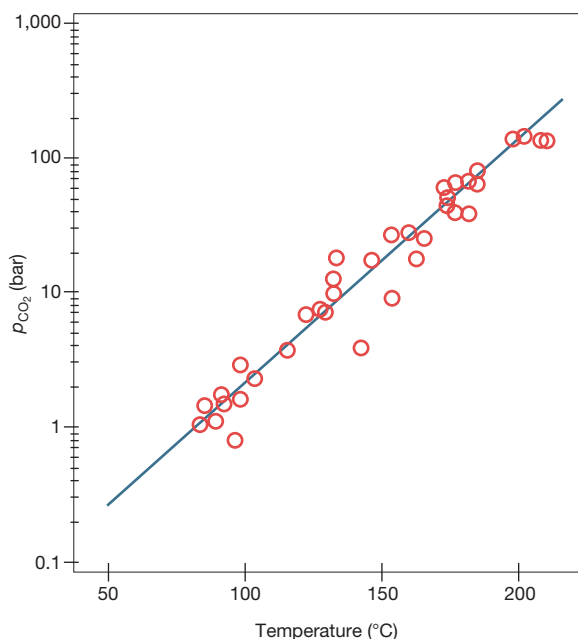


Figure 4 Variations in the partial pressure of carbon dioxide with subsurface temperature in natural gas from the US Gulf Coast (modified from ref. 79). The systematic increase in p_{CO_2} with increasing temperature is consistent with aluminosilicate buffering of carbon dioxide abundance in natural gas.

decarboxylation and oxidation^{51–53} allows their accumulation to high concentration in aqueous fluids during oxidation of *n*-alkanes and other petroleum hydrocarbons^{38,48,54,55}.

In addition to anhydrite in evaporite deposits, there are numerous other oxidants present in sedimentary basins that may react with hydrocarbons. Haematite, magnetite, pyrite and iron-bearing aluminosilicates, such as smectite, are ubiquitously present in sedimentary basins and contain iron or sulphur that may oxidize petroleum and generate organic acids and carbon dioxide^{56,57}. Reduction of ferric minerals at temperatures consistent with petroleum generation indicates that iron-bearing minerals are reactive inorganic sedimentary components that can oxidize hydrocarbons. In contrast to TSR, which is easily identified by increased hydrogen sulphide, evidence for oxidative degradation involving iron-bearing minerals may be limited to increased concentrations of oxygenated alteration products such as carboxylic acids and carbon dioxide. The ubiquitous occurrence of carboxylic acids at concentrations as high as 1,000 mmol in oilfield brines that have not experienced extensive TSR is consistent with pervasive hydrocarbon oxidation in petroleum-producing sedimentary basins (ref. 58, Fig. 3a). The isotope composition of carbon in carbonate cements in the US Gulf Coast decreases systematically with increasing temperature to $\delta^{13}C$ values as low as -16‰ at temperatures as high as 200°C ⁵⁸ (Fig. 3b). This trend is consistent with high-temperature oxidation of isotopically depleted organic carbon in petroleum and kerogen to produce carbon dioxide.

The extent to which oxidative degradation influences the bulk composition of petroleum in subsurface environments devoid of sulphate-bearing evaporites is at present unclear. Such an assessment is difficult because the progressive conversion of long-chain hydrocarbons to shorter chain alteration products during oxidative degradation may be indistinguishable from those attributed to similar processes during thermal cracking. The profound effect of TSR on petroleum composition, however, suggests that interaction of

hydrocarbons with rocks containing abundant ferric iron or other oxidants in migration conduits and reservoirs may represent a previously unrecognized process that has the potential to strongly influence the chemical evolution of oil and natural gas.

Secondary porosity

Secondary porosity is the post-depositional enlargement of pore spaces in sedimentary rocks that results from the leaching of framework minerals and carbonate cements. This topic has been the focus of extensive research because it represents a process that can enhance the reservoir properties of a rock. The percentage of porosity in a reservoir that is secondary in origin is highly variable and difficult to estimate, but values near 10% are routinely reported in the literature. Because the solubility of most minerals is strongly dependent on pore fluid pH, weakly acidic carbon dioxide and carboxylic acids derived from the thermal maturation of organic matter have been implicated in many studies as causative agents during the creation of secondary porosity^{56,58–60}. Carboxylic acids may also increase the solubility of aluminosilicate minerals through the formation of aqueous aluminium–organic acid anion complexes⁶¹. Much of the discussion regarding the influence of carbon dioxide and carboxylic acids has centred around the amounts needed to create the observed volumes of secondary porosity compared to the quantities that can be produced from kerogen. Mass balance calculations that assume that the oxygen required for the formation of carbon dioxide and carboxylic acids is derived solely from kerogen indicate that insufficient quantities of these chemical species can be generated to account for the quantities of secondary porosity seen in the Texas Gulf Coast and Denver Basin^{62–64}.

Reaction of generated hydrocarbons and kerogen with ferric-iron-bearing minerals represents a mechanism to increase yields of carboxylic acids and carbon dioxide beyond the amounts constrained by kerogen–oxygen content^{56,57}. Results of laboratory experiments suggest that reaction pathways exist for the oxidation of hydrocarbons to carboxylic acids and carbon dioxide. In the absence of mineral oxidants, hydrolytic disproportionation may also represent an effective means of generating acidic alteration products. Accordingly, mass-balance constraints based on kerogen–oxygen content may underestimate the amount of carbon dioxide and carboxylic acids generated in sedimentary basins.

The chemical interactions of petroleum with water and minerals can result in the generation of carbon dioxide and organic acids in reservoirs. Relative to the formation of carbon dioxide and carboxylic acids from kerogen in source rocks, *in situ* reservoir production is advantageous for the creation of secondary porosity; this is because it avoids the possibility that the acidity of these highly reactive species may be titrated by minerals in source rocks and migration conduits before reaching the reservoir⁶⁵. Thus, petroleum accumulation may create a corrosive environment that increases porosity through extensive inorganic–organic interactions^{57,66}. Extensive bleaching of red (haematite-rich) sandstones in reservoir rocks suggests that such a process occurs in natural environments⁵⁷.

The ability of carbon dioxide and organic acids to dissolve sedimentary rocks is further complicated by chemical reactions involving aluminosilicates. Increased dissolution of carbonate cements in response to increased abundances of carbon dioxide will only occur if pore fluid pH is not buffered by chemical reactions other than carbonate equilibria. Thermodynamic modelling has demonstrated that pH is buffered by aluminosilicate assemblages in many sedimentary basins at temperatures associated with petroleum generation and accumulation⁶⁷. Decoupling fluid pH from carbonate equilibria suggests that increasing the partial pressure of carbon dioxide in these environments will result in precipitation of carbonate cements instead of dissolution²⁵. Furthermore, the ability of carboxylic acidity to dissolve calcite may be reduced by several orders of magnitude by reaction with aluminosilicate pH buffers⁶⁷. Clearly, our understanding of the processes responsible for the development of secondary

porosity is incomplete, and future efforts in this direction need to account for pH-dependent aluminosilicate reactions as well as the corrosive effects of organically derived carbon dioxide and carboxylic acids.

Natural gas

A high degree of compositional heterogeneity is observed in natural gas deposits owing to variations in the relative abundance of low-molecular-mass hydrocarbons and non-hydrocarbons such as carbon dioxide, hydrogen sulphide and nitrogen. Natural gas is classified as dry if the hydrocarbon component is composed predominantly of methane; wet gas contains significant quantities of C_{2+} hydrocarbons. Although large quantities of methane are produced through the activity of microbial communities living in relatively shallow sediments, the majority of economic natural gas deposits are generated from the thermal alteration of kerogen (primary gas) and oil (secondary gas). Primary gas produced during mainstage oil generation typically contains significant quantities of C_{2+} . Secondary gases may initially be wet, but generally become increasingly methane-rich with increasing thermal maturity.

Despite being the subject of extensive investigation, the origin of dry thermogenic natural gas remains one of the most controversial issues in petroleum geochemistry. Conventional wisdom holds that natural gas deposits are created by thermal cracking of C_{2+} hydrocarbons in oil and wet gas in reservoirs^{1,2,68,69}. Arguments against such a model are based on results of laboratory experiments that indicate that thermal cracking of long-chain hydrocarbons produces a gas enriched in C_2 – C_4 hydrocarbons, and rates of C_2 – C_4 cracking are prohibitively slow at temperatures typical of most oil reservoirs^{70–72}. Evidence that suggests that thermal cracking of hydrocarbons does not occur in nature is provided by the persistence of hydrocarbons in geological environments at temperatures greater than thermal cracking models would predict^{18,72,73}. In response to these issues, numerous alternative hypotheses to account for the generation of methane-rich thermogenic natural gas have been developed. Unequivocal evidence to support a single model is at present not available, and all models may be partly correct.

Phase partitioning in reservoirs and fractionation during migration might be important processes in the formation of methane-rich natural gas deposits⁷⁴. Partitioning–fractionation models are based on the observation that natural gas in source rocks contains substantial quantities of C_{2+} hydrocarbons, and only becomes enriched after upward migration into reservoirs. Alternative models advocate demethylation of aromatic ring systems during heating of residual kerogen deep within sedimentary basins as a mechanism for the generation of dry gas⁷².

Transition-metal catalysis may circumvent issues associated with sluggish reaction kinetics for thermal cracking of hydrocarbons for the conversion of oil to dry natural gas^{70,71,75}. Although direct evidence for transition-metal catalysis in natural environments is not available, decomposition products of long-chain alkenes in the presence of a petroleum source rock heated at 200 °C produces gaseous products that resemble naturally generated gas⁷⁰. Metal porphyrins and metals adsorbed on clays are possible sources for catalytically active transition metals⁷⁵.

The high reactivity of *n*-alkanes dissolved in water suggests that aqueous oxidation or hydrolytic disproportionation reactions play an active part in the conversion of oil to dry gas. Aqueous oxidation, for example, represents an effective mechanism to remove the C_{2+} component of wet gas and to generate methane. Whereas the calculated half-life for thermal cracking of propane at 300 °C is 3,300 yr (ref. 76), the half-life for propane oxidation in the presence of a haematite–magnetite–pyrite mineral assemblage at the same temperature is approximately 2 yr (ref. 38).

Participation of water in processes associated with the generation of oil and natural gas results in the formation of carbon dioxide as a major, if not dominant, alteration product. Most natural gas accu-

mulations contain less than 10% carbon dioxide^{58,63,77–79}, which, at first glance, appears inconsistent with a significant role for water. Partial pressures of carbon dioxide that increase systematically with increasing temperature in sedimentary basins (Fig. 4), however, are consistent with values predicted for carbon dioxide buffering by feldspar, clay and carbonate minerals^{67,79,80}, and suggest that organically derived carbon dioxide may be removed from natural gas by mineral precipitation. The existence of $\delta^{13}C$ -depleted carbon dioxide⁵⁸ and carbonate cements at deep levels in sedimentary basins⁷⁹ (Fig. 3b) indicates that organically derived carbon dioxide is produced from hydrolytic disproportionation or oxidation reactions, but is subsequently removed by mineral precipitation.

Summary and implications

Recognition that water and minerals are reactive sources of hydrogen and oxygen has profound implications for models used to predict the generation of organic alteration products in sedimentary basins. The long-held assumption that hydrogen and oxygen in petroleum are derived solely from kerogen, in conjunction with compositional variations in kerogen as a function of increasing thermal maturity, has been used to constrain the composition, amount and timing of petroleum generation^{1,2,68,69,81}. Examination of Fig. 1b reveals that traditional models predict generation of carbon dioxide and organic acids before mainstage petroleum generation, because the early stages of kerogen maturation result in a substantial decrease in kerogen–oxygen content (Fig. 1a). Similarly, the area under the methane-generation curve in Figure 1b is limited by the hydrogen content of the initial kerogen. Derivation of hydrogen and oxygen from water and minerals eliminates these constraints and suggests that quantities of oxygen- and hydrogen-bearing alteration products may be substantially greater (Fig. 1c). Moreover, generation of carbon dioxide and organic acids may occur before, during and after mainstage petroleum generation, consistent with the occurrence of these species at more than 160 °C (Fig. 3).

Inclusion of water-derived hydrogen in models for the prediction of hydrocarbon generation will affect the estimated quantities of natural gas to a greater extent than similar estimates for oil. Although water-derived hydrogen may play a central part in the reactions responsible for the primary generation of oil, the net hydrogen demand during cleavage of hydrogen-saturated kerogen fragments that constitute oil is relatively minor. By contrast, conversion of long-chain saturated hydrocarbons, with hydrogen:carbon ratios approaching two, to methane with a hydrogen:carbon ratio of four requires the addition of substantial amounts of hydrogen. Because natural-gas generation in the presence of water may only be limited by the availability of a reactive carbon source, models constrained by the compositional evolution of kerogen may significantly underestimate the amount of natural gas, especially in deep basin environments, where mature kerogen is low in hydrogen.

Theoretical, experimental and field studies have provided compelling evidence that the presence of water, minerals and catalytically active transition metals in sedimentary basins has the potential to substantially influence the generation and accumulation of petroleum. Accordingly, conventional models that view the chemical evolution of petroleum as a kinetic process controlled by time, temperature and the composition of initial kerogen alone may not fully account for geochemical processes that result in the formation of petroleum deposits. □

doi: 10.1038/nature02132

1. Tissot, B. P. & Welte, D. H. *Petroleum Formation and Occurrence* (Springer, New York, 1984).
2. Hunt, J. M. *Petroleum Geochemistry and Geology* (W.H. Freeman, San Francisco, 1996).
3. Miknis, F. P., Turner, T. E., Berdan, G. L. & Conan, P. J. Formation of soluble products from thermal decomposition of Colorado and Kentucky oil shales. *Energy Fuels* **1**, 477–483 (1987).
4. Lewan, M. D. Experiments on the role of water in petroleum formation. *Geochim. Cosmochim. Acta* **61**, 3691–3723 (1997).
5. Ziegler, E. R. & Gorman, J. W. Kinetic modelling with multiresponse data. *Technometrics* **22**, 139–151 (1980).
6. Burnham, A. K. in *Composition, Geochemistry and Conversion of Oil Shales* (ed. Shape, C. E.) 211–227

- (D. Reidel, Dordrecht, 1995).
7. Kissin, Y. V. Catagenesis and composition of petroleum: origin of *n*-alkanes and isoalkanes in petroleum. *Geochim. Cosmochim. Acta* **51**, 2445–2457 (1987).
 8. Hunt, J. M. Generation and migration of light hydrocarbons. *Science* **226**, 1265–1270 (1984).
 9. Lewan, M. D. Evaluation of petroleum generation by hydrous pyrolysis. *Phil. Trans. R. Soc. Lond.* **315**, 123–134 (1985).
 10. Orr, W. L. Kerogen/asphaltene/sulfur relationships in sulfur-rich Monterey oils. *Org. Geochem.* **10**, 499–516 (1986).
 11. Baskin, D. K. & Peters, K. E. Early generation characteristics of a sulfur-rich Monterey oils. *Org. Geochem.* **10**, 499–516 (1986).
 12. Tomic, J., Behar, F., Vandenbroucke, M. & Tang, Y. Artificial maturation of Monterey kerogen (Type II-S) in a closed system and comparison with Type II kerogen: implications on the fate of sulfur. *Org. Geochem.* **23**, 647–660 (1995).
 13. Lewan, M. D. Sulfur-radical control on rates of natural petroleum formation. *Nature* **391**, 164–166 (1998).
 14. Brooks, B. T. Evidence of catalytic action in petroleum formation. *Indust. Eng. Chem.* **44**, 2570–2577 (1952).
 15. Goldstein, T. P. Geocatalytic reactions in formation and maturation of petroleum. *Bull. Am. Assoc. Petrol. Geol.* **67**, 152–159 (1983).
 16. Tannenbaum, E., Huizinga, B. J. & Kaplan, I. R. Role of minerals in thermal alteration of organic matter-II: A material balance. *Bull. Am. Assoc. Petrol. Geol.* **70**, 1156–1165 (1986).
 17. Tannenbaum, E. & Kaplan, I. R. Low-M, hydrocarbons generated during hydrous and dry pyrolysis of kerogen. *Nature* **317**, 708–709 (1985).
 18. Price, L. C. Thermal stability of hydrocarbons in nature: Limits, evidence, characteristics, and possible controls. *Geochim. Cosmochim. Acta* **57**, 3261–3280 (1993).
 19. Hoering, T. C. Thermal reaction of kerogen with added water, heavy water, and pure organic substances. *Org. Geochem.* **5**, 267–278 (1984).
 20. Stalker, L., Farrimond, P. & Larter, S. R. Water as an oxygen source for the production of oxygenated compounds (including CO₂ precursors) during kerogen maturation. *Adv. Org. Geochem.* **22**, 477–486 (1994).
 21. Schimmelfmann, A., Bondon, J. P., Lewan, M. D. & Wintsch, R. P. Experimental controls on D/H and ¹³C/¹²C ratios of kerogen, bitumen and oil during hydrous pyrolysis. *Org. Geochem.* **32**, 1009–1018 (2001).
 22. Cooles, G. P., Mackenzie, A. S. & Parkes, R. J. Non-hydrocarbons of significance in petroleum exploration: volatile fatty acids and non-hydrocarbon gases. *Mineral. Mag.* **51**, 483–493 (1987).
 23. Price, L. C. & Wenger, L. M. The influence of pressure on petroleum generation and maturation as suggested by aqueous pyrolysis. *Org. Geochem.* **19**, 141–159 (1992).
 24. Andresen, B., Thronsen, T., Barth, T. & Bolstad, J. Thermal generation of carbon dioxide and organic acids from different source rocks. *Org. Geochem.* **21**, 1229–1242 (1994).
 25. Seewald, J. S., Benitez-Nelson, B. C. & Whelan, J. K. Laboratory and theoretical constraints on the generation and composition of natural gas. *Geochim. Cosmochim. Acta* **62**, 1599–1617 (1998).
 26. Sweeney, J. J., Braun, R. L., Burnham, A. K., Talukdar, S. & Vallejos, C. Chemical kinetic model of hydrocarbon generation, expulsion, and destruction applied to the Maracaibo Basin, Venezuela. *Bull. Am. Assoc. Petrol. Geol.* **79**, 1515–1532 (1995).
 27. Hunt, J. M., Whelan, J. K., Eglinton, L. B. & Cathles, L. M. III. In *Abnormal Pressures in Hydrocarbon Environments* (eds Law, B. E., Ulmishek, G. F. & Slavin, V. I.) 87–104 (AAPG Memoir 70, 1998).
 28. Espitalie, J., Makadi, K. S. & Trichet, J. Role of mineral matrix during kerogen pyrolysis. *Org. Geochem.* **6**, 365–382 (1984).
 29. Horsfield, B. & Douglas, A. G. The influence of minerals on the pyrolysis of kerogens. *Geochim. Cosmochim. Acta* **44**, 1119–1131 (1980).
 30. Huizinga, B. J., Tannenbaum, E. & Kaplan, I. R. The role of minerals in the thermal alteration of organic matter-III. Generation of bitumen in laboratory experiments. *Org. Geochem.* **11**, 591–604 (1987).
 31. Lafargue, E., Espitalie, J., Jacobsen, T. & Eggen, S. Experimental simulation of hydrocarbon expulsion. *Org. Geochem.* **16**, 121–131 (1989).
 32. Treiber, L. E., Archer, D. L. & Owens, W. W. Laboratory evaluation of the wettability of 55 oil producing reservoirs. *Soc. Petrol. Engin. J.* **12**, 531–540 (1992).
 33. Karlsen, D., Nedkvitne, T., Larter, S. R. & Bjørlykke, K. Hydrocarbon composition of authigenic inclusions: Application to elucidation of petroleum reservoir filling history. *Geochim. Cosmochim. Acta* **57**, 3641–3659 (1993).
 34. Siskin, M. & Katritzky, A. R. Reactivity of organic compounds in hot water: Geochemical and technological implications. *Science* **254**, 231–237 (1991).
 35. Taylor, P., Bennett, B., Jones, M. & Larter, S. The effect of biodegradation and water washing on the occurrence of alkylphenols in crude oils. *Org. Geochem.* **32**, 341–358 (2001).
 36. Seewald, J. S. Evidence for metastable equilibrium between hydrocarbons under hydrothermal conditions. *Nature* **370**, 285–287 (1994).
 37. McCollom, T. M., Seewald, J. S. & Simoneit, B. R. T. Reactivity of monocyclic aromatic compounds under hydrothermal conditions. *Geochim. Cosmochim. Acta* **65**, 455–468 (2001).
 38. Seewald, J. S. Aqueous geochemistry of low molecular weight hydrocarbons at elevated temperatures and pressures: constraints from mineral buffered laboratory experiments. *Geochim. Cosmochim. Acta* **65**, 1641–1644 (2001).
 39. Helgeson, H. C., Knox, A. M., Owens, C. E. & Shock, E. L. Petroleum, oil field waters, and authigenic mineral assemblages: Are they in metastable equilibrium in hydrocarbon reservoirs? *Geochim. Cosmochim. Acta* **57**, 3295–3339 (1993).
 40. Orr, W. L. Changes in sulfur content and isotopic ratios of sulfur during petroleum maturation; study of Big Horn Basin Paleozoic oils, advances in petroleum geochemistry. *Bull. Am. Assoc. Petrol. Geol.* **58**, 2295–2318 (1974).
 41. Krouse, H. R., Viau, C. A., Eliuk, L. S., Ueda, A. & Halas, S. Chemical and isotopic evidence of thermochemical sulfate reduction by light hydrocarbon gases in deep carbonate reservoirs. *Nature* **333**, 415–419 (1988).
 42. Machel, H. G., Krouse, H. R., Riciputi, L. R. & Cole, D. R. in *Geochemical Transformations of Sedimentary Sulfur* (eds Vairavamurthy, M. A. & Schoonen, M. A. A.) 439–454 (Am. Chem. Soc., Washington DC, 1995).
 43. Worden, R. H., Smalley, P. C. & Oxtoby, N. H. Gas souring by thermochemical sulfate reduction at 140°C. *Bull. Am. Assoc. Petrol. Geol.* **79**, 854–863 (1995).
 44. Machel, H. G. Gas souring by thermochemical sulfate reduction at 140°C: Discussion. *Bull. Am. Assoc. Petrol. Geol.* **82**, 1870–1873 (1998).
 45. Sassen, R. Geochemical and carbon isotopic studies of crude oil destruction, bitumen precipitation, and sulfate reduction, in the deep Smackover Formation. *Org. Geochem.* **12**, 351–361 (1988).
 46. Claypool, G. E. & Mancini, E. A. Geochemical relationships of petroleum in Mesozoic reservoirs to carbonate source rocks of Jurassic Smackover Formation, southwestern Alabama. *Bull. Am. Assoc. Petrol. Geol.* **73**, 904–924 (1989).
 47. Manzano, B. K., Fowler, M. G. & Machel, H. G. The influence of thermochemical sulfate reduction on hydrocarbon composition in Nisku reservoirs, Brazeau River area, Alberta, Canada. *Org. Geochem.* **27**, 507–521 (1997).
 48. Toland, W. G., Hagman, D. L., Wilkes, J. B. & Brutschy, F. J. Oxidation of organic compounds with aqueous base and sulfur. *J. Am. Chem. Soc.* **80**, 5423–5427 (1958).
 49. Toland, W. G. Oxidation of organic compounds with aqueous sulfur. *J. Am. Chem. Soc.* **82**, 1911–1916 (1960).
 50. Goldhaber, M. B. & Orr, W. L. in *Geochemical Transformations of Sedimentary Sulfur* (eds Vairavamurthy, M. A. & Schoonen, M. A. A.) 412–425 (Am. Chem. Soc., Washington, DC, 1995).
 51. Bell, J. L. S. & Palmer, D. A. in *Organic Acids in Geological Processes* (eds Pittman, E. D. & Lewan, M. D.) 227–269 (Springer, New York, 1994).
 52. McCollom, T. S. & Seewald, J. S. Experimental constraints on the hydrothermal reactivity of organic acids and acid anions: I. Formic Acid and formate. *Geochim. Cosmochim. Acta* **67**, 3625–3644 (2003).
 53. McCollom, T. S. & Seewald, J. S. Experimental study of the hydrothermal reactivity of organic acids and acid anions: II. Acetic acid, acetate, and valeric acid. *Geochim. Cosmochim. Acta* **67**, 3645–3664 (2003).
 54. Eglinton, T. I., Curtis, C. D. & Rowland, S. J. Generation of water-soluble organic acids from kerogen during hydrous pyrolysis: implications for porosity development. *Mineral. Mag.* **51**, 495–503 (1987).
 55. Borgund, A. E. & Barth, T. Generation of short-chain organic acids from crude oil by hydrous pyrolysis. *Org. Geochem.* **21**, 943–952 (1994).
 56. Surdam, R. C. & Crossey, L. J. Organic-inorganic reactions during progressive burial: key to porosity and permeability enhancement and preservation. *Phil. Trans. R. Soc. Lond.* **315**, 135–156 (1985).
 57. Surdam, R. C., Jiao, Z. S. & MacGowan, D. B. Redox reactions involving hydrocarbons and mineral oxidants: A mechanism for significant porosity enhancement in sandstones. *Bull. Am. Assoc. Petrol. Geol.* **77**, 1509–1518 (1993).
 58. Franks, S. G. & Forester, R. W. in *Clastic Diagenesis* (eds McDonald, D. A. & Surdam, R. C.) 63–80 (Am. Assoc. Petrol. Geol., Tulsa, 1984).
 59. Schmidt, V. & MacDonald, D. A. in *Aspects of Diagenesis*. (eds Scholle, P. A. & Schluger, P. R.) 175–207 (Soc. Econ. Paleontol. Min., Tulsa, 1979).
 60. Surdam, R. C., Crossey, L. J., Hagen, E. S. & Heasler, H. P. Organic-inorganic interactions and sandstone diagenesis. *Bull. Am. Assoc. Petrol. Geol.* **73**, 1–23 (1989).
 61. Crossey, L. J., Surdam, R. C. & Lahann, R. in *Roles of Organic Matter in Sediment Diagenesis* (ed. Gautier, D.) 147–155 (Soc. Econ. Paleont. Min., Tulsa, 1986).
 62. Lundegard, P. D., Land, L. S. & Galloway, W. E. Problem of secondary porosity: Frio Formation (Oligocene), Texas Gulf Coast. *Geology* **12**, 399–402 (1984).
 63. Lundegard, P. D. & Land, L. S. in *Roles of Organic Matter in Sediment Diagenesis* (ed. Gautier, D. L.) 129–146 (Soc. Econ. Paleont. Min., Tulsa, 1986).
 64. Pittman, E. D. & Hathon, L. A. in *Organic Acids in Geological Processes* (eds Pittman, E. D. & Lewan, M. D.) 115–137 (Springer, New York, 1994).
 65. Giles, M. R., de Boer, R. B. & Marshall, J. D. in *Organic Acids in Geological Processes* (eds Pittman, E. D. & Lewan, M. D.) 449–470 (Springer, New York, 1994).
 66. Price, L. C. & DeWitt, E. Evidence and characteristics of hydrolytic disproportionation of organic matter during metasomatic processes. *Geochim. Cosmochim. Acta* **65**, 3791–3826 (2001).
 67. Hutcheon, I., Shevalier, M. & Abercrombie, H. J. pH buffering by metastable mineral-fluid equilibria and evolution of carbon dioxide fugacity during burial diagenesis. *Geochim. Cosmochim. Acta* **57**, 1017–1027 (1993).
 68. Barker, C. Calculated volume and pressure changes during the thermal cracking of oil to gas in reservoirs. *Bull. Am. Assoc. Petrol. Geol.* **74**, 1254–1261 (1990).
 69. Ungerer, P. State of the art research in kinetic modeling of oil formation and expulsion. *Org. Geochem.* **16**, 1–25 (1990).
 70. Mango, F. D., Hightower, J. W. & James, A. T. Catalysis in the origin of natural gas. *Nature* **368**, 536–538 (1994).
 71. Mango, F. D. The light hydrocarbons in petroleum: a critical review. *Org. Geochem.* **26**, 417–440 (1997).
 72. McNeil, R. I. & BeMent, W. O. Thermal stability of hydrocarbons: laboratory criteria and field examples. *Energy Fuels* **10**, 60–67 (1996).
 73. Sugisaki, R. & Mimura, K. Mantle hydrocarbons: Abiotic or biotic? *Geochim. Cosmochim. Acta* **58**, 2527–2542 (1994).
 74. Price, L. C. & Schoell, M. Constraints on the origins of hydrocarbon gas from compositions of gases at their site of origin. *Nature* **378**, 368–371 (1995).
 75. Mango, F. D. The origin of light hydrocarbons. *Geochim. Cosmochim. Acta* **64**, 1265–1277 (2000).
 76. Laidler, K. J., Sagert, N. H. & Wojciechowski, B. W. Kinetics and mechanisms of the thermal decomposition of propane. *Proc. R. Soc. Lond.* **270**, 242–253 (1962).
 77. Jenden, P. D., Kaplan, I. R., Poreda, R. J. & Craig, H. Origin of nitrogen-rich natural gases in the California Great Valley; evidence from helium, carbon and nitrogen isotope ratios. *Geochim. Cosmochim. Acta* **52**, 851–861 (1988).
 78. Jenden, P. D., Newell, K. D., Kaplan, I. R. & Watney, W. L. Composition and stable-isotope geochemistry of natural gases from Kansas, midcontinent, U.S.A. *Chem. Geol.* **71**, 117–147 (1988).
 79. Smith, J. T. & Ehrenberg, S. N. Correlation of carbon dioxide abundance with temperature in clastic hydrocarbon reservoirs: relationship to inorganic chemical equilibrium. *Mar. Petrol. Geol.* **6**, 129–135 (1989).
 80. Hutcheon, I. & Abercrombie, H. Carbon dioxide in clastic rocks and silicate hydrolysis. *Geology* **18**, 541–544 (1990).
 81. Cooles, G. P., Mackenzie, A. S. & Quigley, T. M. Calculation of petroleum masses generated and expelled from source rocks. *Org. Geochem.* **10**, 325–345 (1986).
 82. Shock, E. L. *Organic Acids in Geological Processes* (eds Pittman, E. D. & Lewan, M. D.) 270–318 (Springer, New York, 1994).
 83. O'Neil, J. R., Clayton, R. N. & Mayeda, T. K. Oxygen isotope fractionation in divalent metal carbonates. *J. Chem. Phys.* **51**, 5547–5558 (1969).

Understanding the thermal evolution of deep-water continental margins

Nicky White^{1*}, Mark Thompson² & Tony Barwise²

¹Bullard Laboratories, Department of Earth Sciences, Madingley Rise, Madingley Road, Cambridge, CB3 0EZ, UK

(e-mail: nwhite@esc.cam.ac.uk)

²BP Exploration Operating Company Ltd, Compass Point, 79–87 Kingston Road, Knowle Green, Staines, Middlesex, TW18 1DY, UK

*Present address: Department of Geology, Trinity College, Dublin 2, Ireland

Areas of exploration for new hydrocarbons are changing as the hydrocarbon industry seeks new resources for economic and political reasons. Attention has turned from easily accessible onshore regions such as the Middle East to offshore continental shelves. Over the past ten years, there has been a marked shift towards deep-water continental margins (500–2,500 m below sea level). In these more hostile regions, the risk and cost of exploration is higher, but the prize is potentially enormous. The key to these endeavours is a quantitative understanding of the structure and evolution of the thinned crust and lithosphere that underlie these margins.

Whatever the controversies surrounding our dependency upon fossil fuels, one issue is very clear. There is a finite amount of hydrocarbon left to discover. On the basis of current reserves, liquid hydrocarbon production will peak at ~30 billion barrels per year in 15–20 years (1 barrel contains 0.16 m³ oil), declining to 5–10 billion barrels per year by the end of the twenty-first century^{1,2}. At present, the hydrocarbon industry depends for its economic survival upon its ability to locate and extract hydrocarbons. During the first half of the twentieth century, sites of hydrocarbon production were predominantly located onshore. In the late twentieth century, exploration moved offshore, where drilling costs and exploration risks are potentially much higher. The bulk of major hydrocarbon fields located in shallow-water depths (that is, up to 200 m) have probably been located, if one excludes fields at depths greater than ~5 km and the unexplored margins of Antarctica and the Arctic Ocean.

The hydrocarbon industry is now pursuing two different strategies. The first is to efficiently extract greater amounts of hydrocarbons from existing reserves, from which typically only 30–40% of oil is recovered. This strategy is an obvious one because modest efficiencies in a region with substantial undeveloped reserves will surpass the fruits of exploration elsewhere. The major technical problem concerns the connectivity of reservoir rocks. Its solution relies on a combination of sophisticated engineering (for example, drilling deviated wells to target hydrocarbon ‘sweet-spots’) and time-lapse imaging (that is, repeated seismic monitoring of producing fields where changes in acoustic response enable the movements of subsurface fluids to be tracked). In essence, it is an intimate combination of geophysics, reservoir engineering and economic practicality.

The second strategy is to explore virgin areas, an inherently risky and expensive enterprise. On Earth, the largest piece of unexplored continent consists of submerged continental margins that form a ‘fringe’ around the deep ocean basins (Fig. 1a). Over the past ten years, there has been a relentless drive to explore ever-increasing water depths. This drive has been stimulated by an engineering technology that has allowed us to drill, in water depths of nearly 3 km, 20 cm diameter holes into rocks that lie ~4 km beneath the

seabed (Fig. 1b). From a strictly economic perspective, this strategy is irrational: a single deep-water exploration well costs up to US\$50 million, which is about two orders of magnitude greater than a typical onshore well in the Middle East. However, political and economic realities conspire to make deep-water exploration a commercially viable proposition.

Here, we describe the geological framework that is used to identify the most prolific margins and to guide the search for deep-water hydrocarbons. We are especially concerned with specific criteria that make any margin more attractive for exploration. Our focus, for two reasons, is on the structure of margins that surround the Atlantic Ocean. First, these extensional margins are the most important locations, apart from Southeast Asia, for current deep-water exploration. Second, our understanding of the structure and evolution of extensional continental margins is primarily based upon detailed studies of sediment-starved margins in the North Atlantic Ocean. We exclude oceanic basins and plateaux but include the transition from thinned continental to oceanic crust, which is a future exploration target. Our main theme is the discovery of new hydrocarbon fields rather than the problem of extraction. We do not discuss the economics of drilling and production — suffice it to say that if one can risk substantial capital sums, the rewards are enormous in the event of success.

Identifying prolific margins

Exploration geologists have cherry-picked deep-water margins by identifying those where there is an established hydrocarbon system in shallower water or onshore. There are four prerequisites for a high-quality hydrocarbon system. First, does the margin have world-class source rock that has undergone an appropriate degree of thermal maturation? The distribution of suitable organic-rich deposits is strongly dependent upon the development of basins with restricted circulation and upon the location of coastal upwelling zones in ancient times³ (Fig. 1a). Second, was an adequate supply of sand-rich or carbonate-rich sediment deposited to form suitably porous reservoir rocks? Rivers with large drainage catchments guarantee a substantial input of sand-rich sediment over protracted periods of time (that is, tens of millions of years)⁴. Density underflows

transport these sand-rich sediments down the slope onto the abyssal plains⁵. Third, has the margin been deformed to generate large sub-surface structures suitable for trapping billions of barrels of hydrocarbons? Shallow-water exploration has demonstrated the importance of subtler stratigraphic traps, but during the early stages of exploration, large and easily identifiable structures are favoured. Finally, hydrocarbons are less dense than interstitial waters, and so evidence for an unruptured seal, which would prevent buoyant hydrocarbons from leaking out, is needed. The best seals are fine-grained rocks with low permeabilities (for example, mudstones and evaporites).

These four prerequisites limit us to a small set of regions, which currently include the Gulf of Mexico and the West Africa Margin (Fig. 2). In both cases, there is a favourable collocation of high-quality source rock, high sediment input and numerous structures generated by salt or shale deformation at depth. Both areas have a track record of successful onshore and shallow-water exploration. Their importance is emphasized by reserve and investment statistical analyses⁶. In the past five years, global deep-water reserves totalling ~10 billion barrels of oil-equivalent were brought to production, and this quantity is expected to treble in the next five years. Outside Europe, the estimated reserves located offshore of Central and South America and offshore of Africa dominate, especially in water depths greater than 500 m. The combined capital expenditure for the Gulf of Mexico and West Africa Margins reached US\$4.5 billion in 2002, and is expected to increase rapidly over the next five years. Offshore of West Africa, drilling expenditure in water depths of greater than 500 m now exceeds that spent closer to shore. For the biggest multinational companies, the immediate future for exploration is focused on the Gulf of Mexico and West Africa Margin. These pickings are, in a sense, relatively easy, and a major question is how many unexplored deep-water margins will prove to be as productive.

When the above prerequisites are in place and when the engineering technology is available, deep-water exploration is an attractive proposition. With increasing water depth and escalating drilling costs, two of the most significant uncertainties are the thermal evolution of the continental margin, which determines the maturation history of a given source rock, and the timing of trap formation. Both uncertainties can be minimized by developing integrated thermal and structural models of margin growth.

Formation of continental margins

Over the past 30 years, there has been considerable interest in the evolution of passive continental margins, which are a primary manifestation of plate tectonics. This research was originally motivated by an interest in explaining the pattern of subsidence recorded by deep boreholes⁷. It is now generally accepted that these margins form by extension of the lithospheric plate, which is 120 ± 20 km thick (Fig. 3). Simple kinematic models were developed in the late 1970s and 1980s that describe the growth of margins in terms of β , the stretching factor^{8,9}. The amount and rate of stretching determine the temporal and spatial variation of crustal and lithospheric thinning across the margin, which in turn controls variations in heat flow, subsidence, extensional faulting and decompression melting. During thinning, the base of the lithosphere is advected upwards and heat flow to the surface rapidly increases. At the same time, hotter asthenosphere is passively moved upwards, and generates a thermal anomaly. Within the upper crust, thinning is shown by rapid, fault-controlled subsidence. In general, thinning steadily increases over distances of 100–500 km, reaching ratios of three to four, at which point substantial decompression melting occurs and the oceanic crust is generated at a mid-oceanic ridge system. When stretching stops, the thermal anomaly decays exponentially with time, heat flow decreases and a phase of thermally driven subsidence occurs.

The temporal variation of heat flow directly depends upon the rate of stretching, but it can be moderated in four significant ways^{10–12}. First, heat-producing elements concentrated in the crust

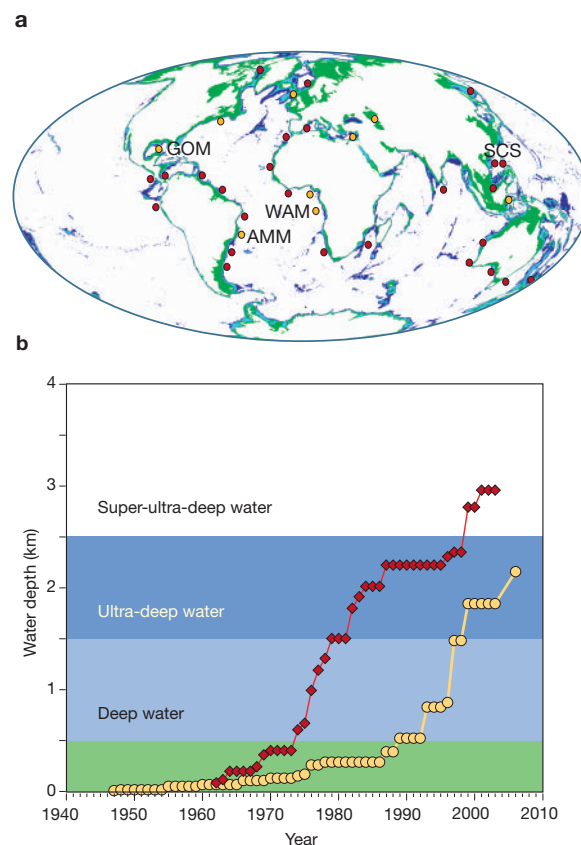


Figure 1 Exploration at deep-water margins. **a**, Hammer Equal Area Projection of the World, which shows the distribution of deep-water margins of interest to the hydrocarbon industry. Green fringes indicate water depths of 0–500 m, light-blue fringes 500–1,500 m and dark blue fringes 1,500–2,500 m. Solid yellow circles represent current deep-water areas, and solid red circles are frontier deep-water areas. Note the level of activity throughout the South Atlantic Ocean. GOM, Gulf of Mexico; WAM, West Africa Margin; AMM, Amazon Margin; SCS, South China Sea. **b**, Worldwide progress in drilling for, and producing, hydrocarbons as a function of water depth, using the colour scheme of **a**. Red diamonds show exploration capability (record of 2,965 m in 2001 held by Transocean Sedco Forex operated by Unocal); yellow circles show production platform/floater systems (record of 1,853 m in 1999 held by Rocader Field operated by Petrobras, a Brazilian company that has been an important pioneer in deep-water exploration; in 2006, anticipated record of 2,146 m for Atlantis Field operated by BP). Note that the lag time between exploration and production drilling is rapidly closing. Some figures were produced using GMT³⁵.

make a contribution that decreases as the crust progressively thins. Second, if large amounts of decompression melting occur and significant volumes of hot magma are situated within the crust, heat flow will increase, albeit for a short period of time. Third, heat flow varies when the crust and lithospheric mantle thin at different rates and by different amounts (Fig. 3d–f). Finally, rapid deposition of cold sediment can perturb the geothermal gradient for short periods of time (1–10 Myr).

Imaging continental margins

Seismic experiments are used to produce images of the general structure of margins. These experiments use acoustic energy generated by the release of compressed air from arrays of large airguns that are towed behind a ship. A small fraction of this energy is transmitted through the Earth and reflected or refracted at interfaces between different rock types, where density and acoustic velocity abruptly change. Energy reflected at small angles of incidence (less than 30°) is

recorded by hydrophones located at intervals on a long (~6–12 km) streamer towed in a straight line behind the ship. Energy reflected and refracted at greater angles can be recorded by seismometers placed at the bottom of the ocean. Travel times and amplitudes of the recorded signals are used to calculate the variation of acoustic velocity down to a depth of ~40 km (that is, the crust and uppermost mantle but not the entire lithospheric plate). An empirical relationship based upon laboratory experiments is used to convert acoustic velocity into density.

These logistically complex and expensive experiments (US\$0.5 million–US\$1 million) yield excellent acoustic images of the crust from the coastline to the abyssal plain where bona fide oceanic crust can be identified. Unfortunately, the highly sedimented and deformed margins favoured by the oil industry are more difficult to image because high frequencies are rapidly attenuated by low-velocity sediments and scattered by complexly deformed strata. The best images are from relatively sediment-starved margins located in the North Atlantic Ocean (Figs 2, 3). There is considerable variation in the detailed structure and symmetry of different margins, but from a thermal perspective we can divide margins into three broad categories.

At one end of the spectrum are 'hot' margins, where lithospheric thinning has taken place over a upwelling mantle plume. The best-studied examples occur in the North Atlantic Ocean on either side of the Iceland Plume (Fig. 3a) (refs 13, 14). These 'hot' margins have two defining characteristics. Close to the seabed, deep-sea drilling has confirmed the existence of kilometre-thick piles of seaward-dipping lava flows. At the base of the thinned crust, large prisms of material with velocities of 7.2–7.6 km s⁻¹ have been imaged. Velocities and inferred densities are consistent with magnesium-rich igneous rocks generated by large-scale decompression melting of asthenosphere during continental break-up. 'Hot' margins present particular difficulties for deep-water exploration. The presence of high-velocity lava

flows, which scatter acoustic energy, considerably impedes our ability to image underlying sedimentary strata. It is also difficult to accurately constrain thermal histories because the spatial and temporal distribution of hot molten rock, which advects heat, is not easy to determine with accuracy. Nonetheless, deep-water exploration of the northwest European continental shelf has met with some success¹⁵.

At the other end of the spectrum are 'cold' margins, where there is little evidence for magmatism until new oceanic crust has formed¹⁶. The best-studied margin occurs west of the Iberian Peninsula¹⁷, but this type of margin is more widespread than 'hot' ones (Figs 2, 3c). As before, crustal thickness decreases from 30 km to less than 5 km over a distance of ~200 km. Within this region, there is excellent evidence from deep-sea drilling that rocks from the mantle were exposed at the sea floor during lithospheric thinning¹⁷. The existence of 'transition zones' of exhumed mantle, which are several tens of kilometres wide was entirely unexpected. These 'transition zones' are characterized by a velocity structure that is different from both the oceanic crust and the stretched continental crust (Fig. 3c). Acoustic velocities rise steeply to ~7 km s⁻¹ only 2 km beneath deformed crustal rocks, increasing gradually back to normal mantle values of ~8 km s⁻¹ over a depth range of 6 km. This velocity variation may be caused by serpentinization of mantle rocks (that is, hydration of iron- and magnesium-rich minerals) as a result of contact with sea water, although there are other possible explanations¹⁸. Within the upper crust, faults flatten significantly with depth and merge with an undulating 'detachment surface' that separates the crust and mantle. This unusual combination of negligible magmatism, exhumed mantle and detachment surfaces is controversial and puzzling. The major difficulty is that exhumation of the mantle should generate significant decompression melting¹⁹. One possible explanation is that lithospheric stretching may have occurred extremely slowly during the early stages of margin formation. Alternatively, negligible

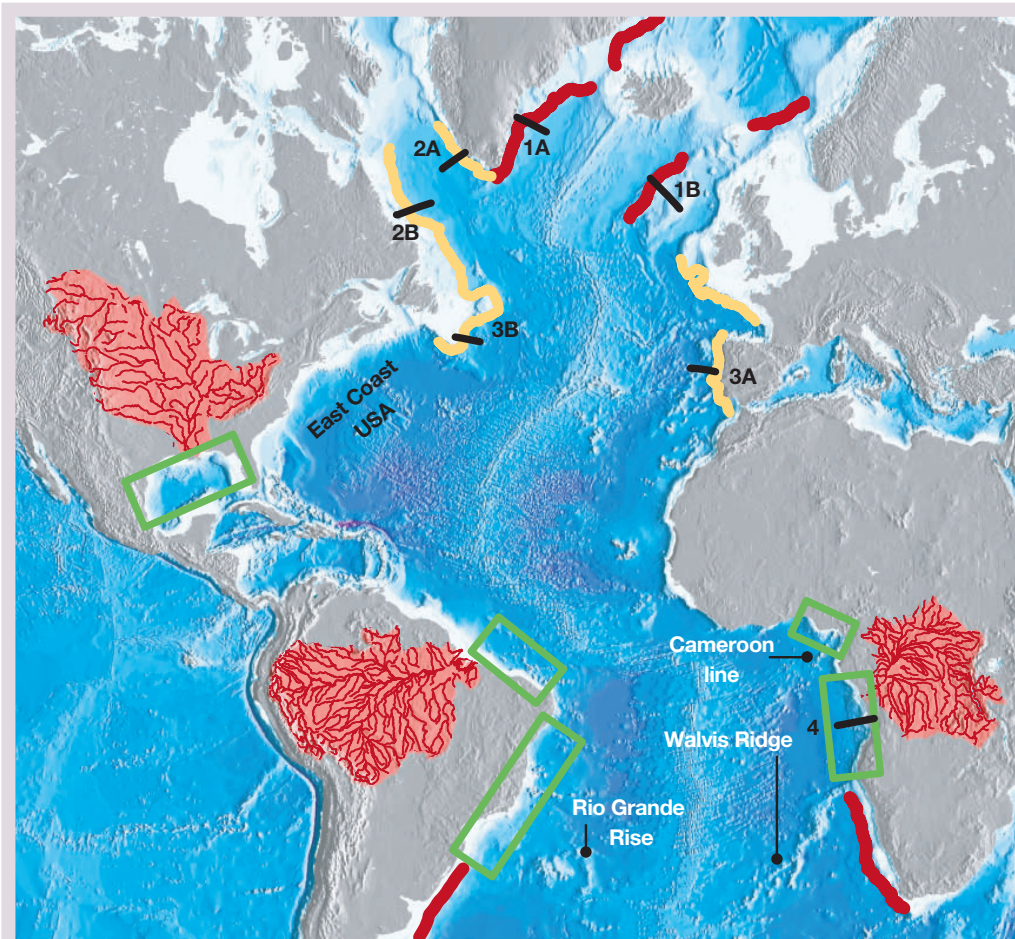


Figure 2 Margin categories in the Atlantic Ocean. Topography and bathymetry of the region that includes the Atlantic Ocean. In the North Atlantic Ocean, locations of three pairs of combined seismic wide-angle and deep reflection profiles are shown: 1A, B, SIGMA 3 and Hatton Bank surveys^{13,14}, 2A, B, Labrador-Greenland survey²⁰, 3A, B, Newfoundland and Iberian margin surveys^{16,17}. 4 indicates the cross-section shown in Fig. 6. Red lines show 'hot' margins, where large volumes of magma were generated by rifting over a mantle plume. Yellow lines show 'cold' margins where there is evidence for exhumation and serpentinization of the lithospheric mantle. The east coast margin of North America is more ambiguous (that is, evidence for some magmatism but no obvious mantle plume). In the central and south Atlantic Ocean, the drainage catchments of the Mississippi, Amazon and Congo rivers are shown as pink and red regions. Two green boxes illustrate zones of deep-water exploration in the Gulf of Mexico and offshore of West Africa, where high-quality source rocks, excellent reservoir sands and salt-related structures coexist. Other green boxes illustrate zones of interest elsewhere, including the Amazon delta, which has potential for oil production.

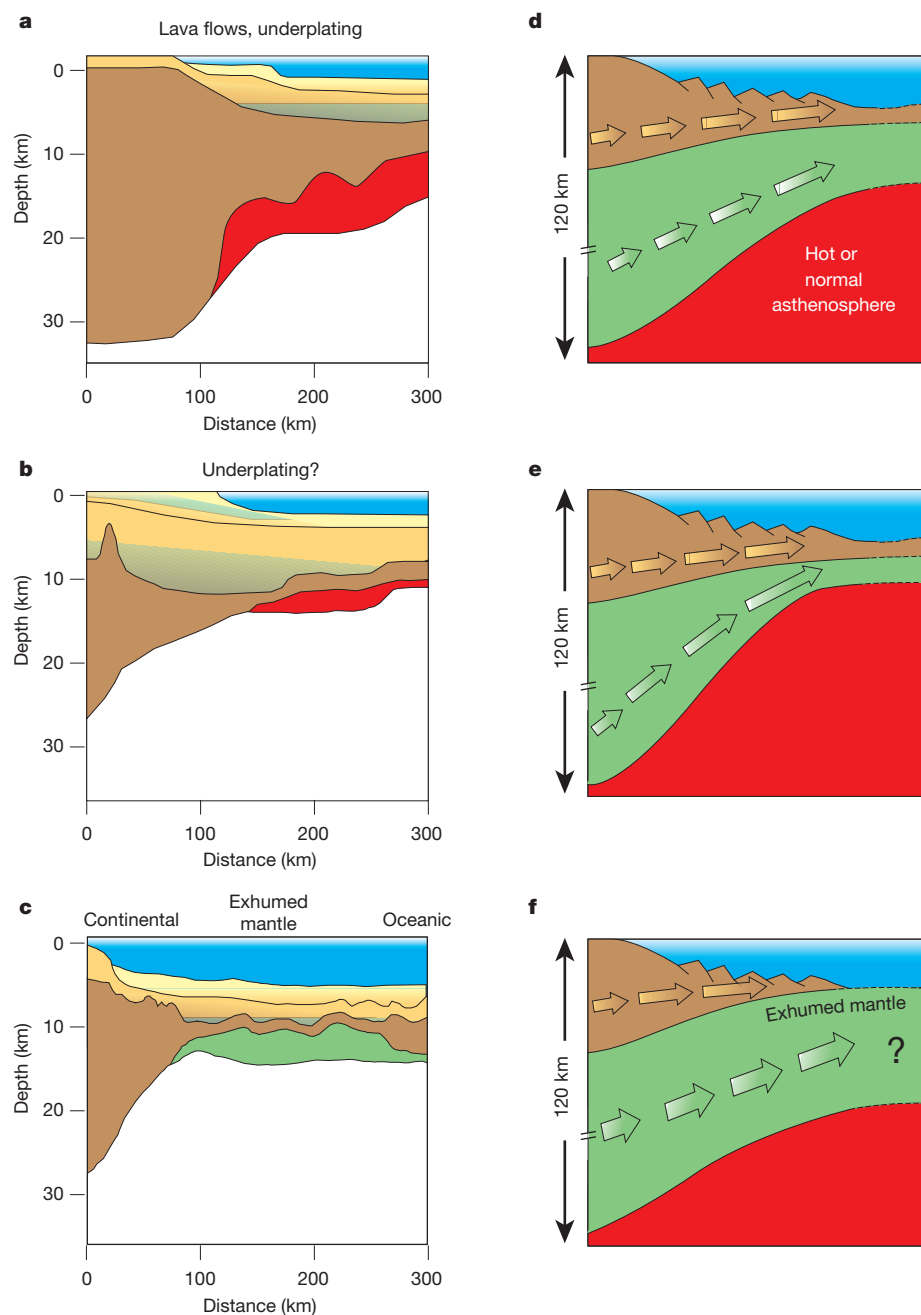


Figure 3 Structure and evolution of extensional margins. **a–c**, Cross-sections that illustrate the crustal structure of one side of three conjugate margin pairs from the North Atlantic Ocean (see Fig. 2 for location). In each case, crustal velocities and boundaries were determined by forward and inverse modelling of wide-angle seismic data in conjunction with deep seismic reflection data and gravity measurements. Blue represents sea water; yellow and light brown represent sedimentary rocks; dark brown represents the crust, red the magmatic underplating and green the serpentinized mantle rock. **a**, Structure of the East Greenland margin from the SIGMA 3 experiment^{13,14}; **b**, structure of the Labrador Sea margin from the 90R1 experiment²⁰; **c**, structure of the Iberian Sea margin^{16,17}. **d–f**, Cartoons illustrating three possible configurations of crustal and lithospheric mantle thinning. **d**, Uniform thinning; **e**, non-uniform thinning with greater lithospheric mantle thinning at the right-hand end; **f**, non-uniform thinning with greater crustal thinning at the right-hand side. The question mark indicates the need for the lithospheric mantle to extend at some point to generate the mid-oceanic ridge system. In each case, the integrated amount of crust and lithospheric mantle extension must balance, but the temporal and spatial patterns of heat flow and subsidence differ. Blue, sea water; brown, crust; green, lithospheric mantle; red, asthenospheric mantle.

stretching of the lithospheric mantle occurred until the oceanic crust was generated (Fig. 3c). The way in which these transition zones form has important implications because the temporal and spatial variation of heat flow will be strongly affected. The conductivity of serpentine is half that of typical mantle minerals.

A third category of margin lies between these two extremes (Fig. 3b)¹⁶. At these margins, there is no convincing evidence in favour of a mantle plume, even though high velocities consistent with magmatic underplating occur in the lower crust¹⁶. The most likely explanation is that a non-plume process generated modest volumes of igneous rock, although it has been suggested that serpentinization of mantle rock plays a part. These 'warm' margins occur along the east coast of North America but are probably more widespread^{20,21}.

In the south Atlantic Ocean, not many combined wide-angle and deep-reflection experiments have been carried out. In many places, existing deep seismic imaging is hampered by thick layers of sediment

and salt. One experiment was carried out on the West Africa Margin 500 km south of the Walvis Ridge that demonstrates that this margin is relatively narrow and strongly influenced by magmatism²². Thick wedges of seaward-dipping lava flows occur beneath the seabed, and a prism of high-velocity material within the lower crust is consistent with magmatic underplating. The conjugate pair of 'hot' margins, which occurs south of the Walvis Ridge and south of the Rio Grande Rise, was formed on top of the Tristan da Cunha Plume during break-up of the south Atlantic Ocean¹³ (Fig. 2). Further north, where the Congo and Amazon rivers debouch, gravity modelling²³ and unpublished seismic wide-angle data suggest that negligible magmatism took place during lithospheric stretching.

How do margins grow?

In regions of active and distributed deformation, earthquake focal mechanism solutions, Global Positioning System (GPS) measure-

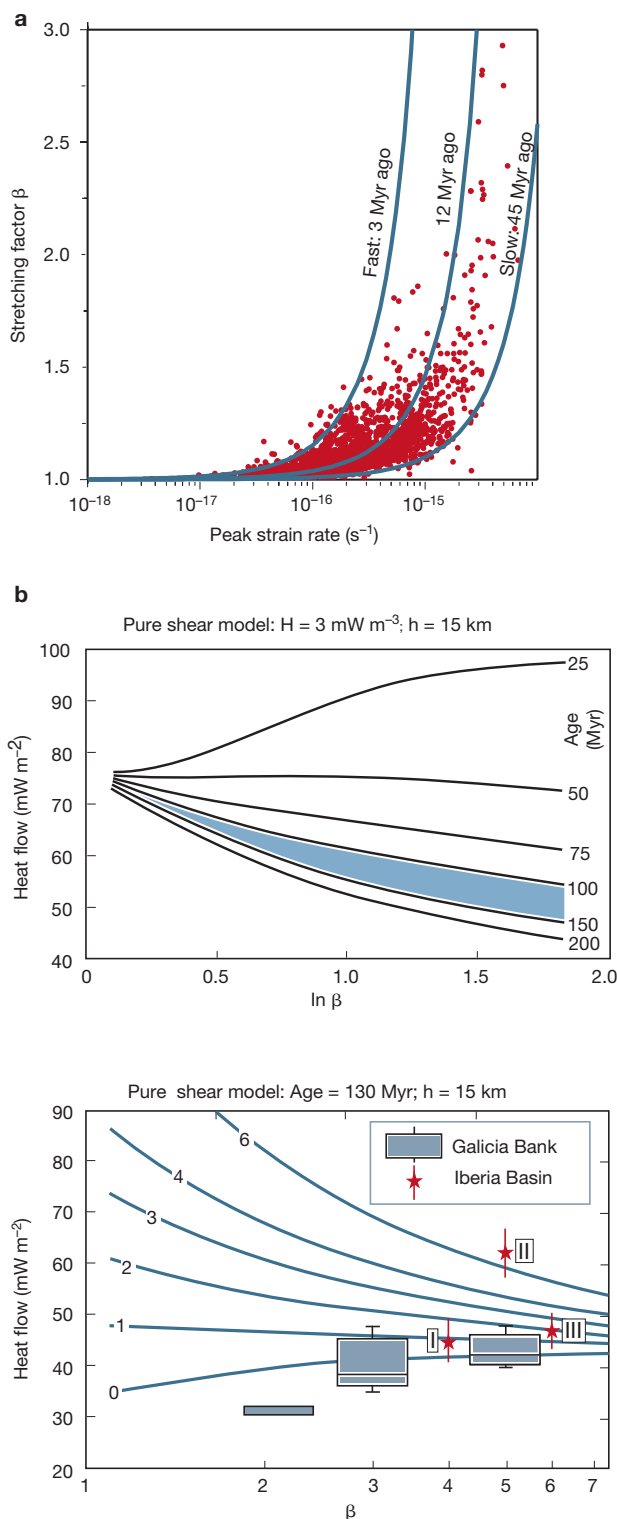


Figure 4 Strain rate and heat flow at margins. **a**, Total stretching factors plotted as a function of peak strain rates. Solid red circles represent values obtained by inverting subsidence data from sedimentary basins and margins located worldwide and where total extension is small ($\beta < 3$) (ref. 27); labelled and numbered blue lines represent variation of the stretching factor as a function of strain rate for given rifting durations that range from fast to slow (that is, 3–40 Myr). **b**, Diagrams that show variation of heat flow as a function of stretching factor and radiogenic crustal components. Red stars and blue boxes are heat-flow measurements (adapted from ref. 11).

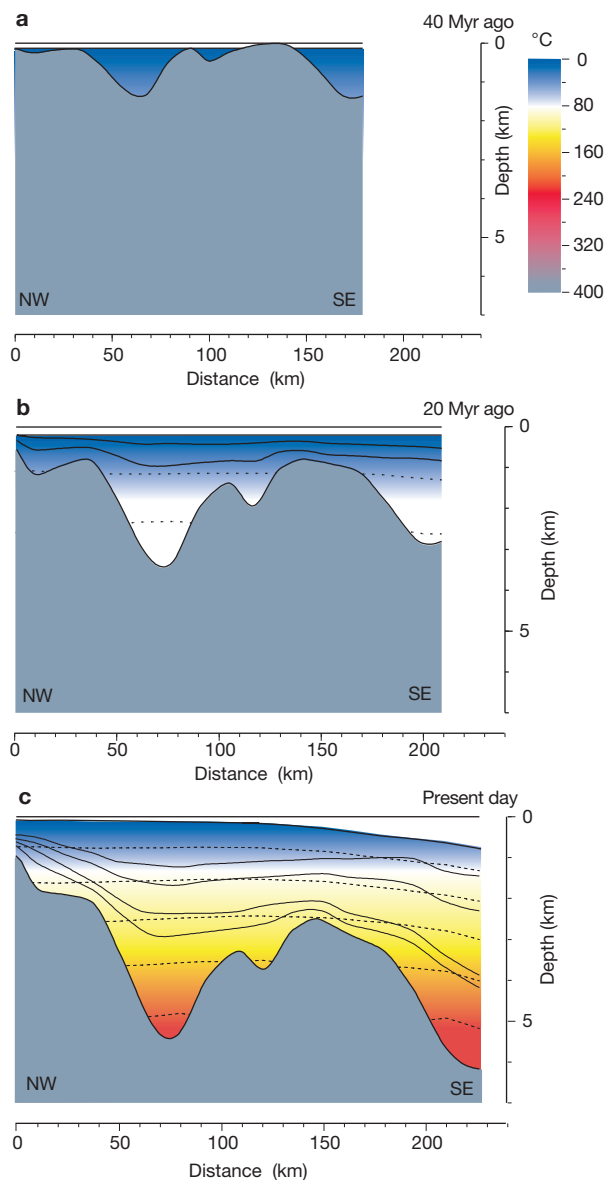


Figure 5 Animated margin evolution. Thermal and structural evolution of the northern margin of the South China Sea. Two-dimensional subsidence data were first inverted to determine the spatial and temporal variation of strain rate. Strain rates were then used to calculate heat-flow variation. If we assume that the growing sedimentary pile has a thermal conductivity structure that is a simple function of compaction, we can calculate temperature through time and space. It is straightforward to include crustal heat production and sediment blanketing, which have a secondary effect on the thermal evolution of slowly extending basins³⁶. Our results are presented as a series of images at different times that can be assembled to make animations of the thermal and structural growth of margins. These animations are powerful commercial tools because they allow us to assess the timing and development of temperature with respect to the structure (S. Jones *et al.*, personal communication). At any given time, maturation of potential source rocks can be calculated from the temperature structure. **a**, Vertical cross-section of a margin at 40 Myr ago (that is, 20 Myr after start of lithospheric extension); solid black lines indicate the subsidence record; cold/warm colours indicate the temperatures of sediment according to colour scale bar; and grey indicates basement rocks. **b**, Cross-section at 20 Myr ago; comparison of horizontal scales yields amount of lithospheric extension. Dotted lines indicate temperature contours at 40° intervals. **c**, Cross-section at the present day.

Box 1

Margin maths

A continuum approach is used to model stretching of the lithospheric plate through space and time³⁷. Accordingly, we must solve:

$$\frac{DF}{Dt} = LF \quad (1)$$

where F is the deformation gradient tensor and L is the velocity gradient tensor whose elements are:

$$L_{ij} = \frac{\partial v_i}{\partial x_j} \quad (2)$$

v_i are the velocities in the x_i directions where i and j vary from 1 to 3. The expression D/Dt in equation (1) is the substantive or lagrangian derivative, which means that the time derivative is applied to a vector joining one pair of particles. Thus at time t , the deformation of a short line, $\mathbf{p}(t)$, within the continuum is given by:

$$\mathbf{p}(t) = F(t)\mathbf{p}(0) \quad (3)$$

where $\mathbf{p}(0)$ is a vector joining two particles at $t=0$. These equations are valid for any temporally and spatially varying velocity field, and they form the basis of many models that describe finite lithospheric deformation^{26,38}. In actively deforming regions, a similar approach is used to obtain the instantaneous horizontal velocity field by inverting strain-rate (that is, the velocity gradient) data on the basis of earthquake focal mechanisms and GPS measurements²⁴. At ancient continental margins where deformation has ceased, the velocity field can be only indirectly determined. One promising approach uses the subsidence history calculated from seismic images that have been calibrated by borehole data. We acknowledge that strain-rate estimates are strongly affected by uncertainties in ancient water depths across sediment-starved continental margins.

Observed subsidence is compared with predicted subsidence, which is calculated by first solving the thermal structure of the lithosphere, $T(x,y,z,t)$, for some arbitrary velocity field. T is calculated by solving the three-dimensional heat-flow equation with appropriate advective terms:

$$\left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla\right) T = \nabla \cdot (\kappa \cdot \nabla T) \quad (4)$$

Equations (1) and (4) are the cornerstones of an algorithm that calculates the subsidence pattern. Neither can be solved analytically for arbitrary velocity fields and numerical methods are used instead^{30,39}. In a smooth two-dimensional model, we assume that horizontal velocity is constant as a function of depth, and we ignore short-wavelength variations associated with upper crustal faults. By definition, the spatial and temporal variation of the stretching factor is $\beta(x,t) = F_{11}$. F is initially a unit matrix, and so equation (3) reduces to:

$$\frac{\partial \beta}{\partial t} + u \frac{\partial \beta}{\partial x} = \frac{\partial u}{\partial x} \beta \quad (5)$$

The two-dimensional algorithm is divided into four parts. First, the variation of strain rate through space and time is defined and used to calculate the velocity field. Second, this velocity field determines the evolving thermal structure. Third, the changing density structure, which defines the history of lithospheric loading, is calculated from the thermal structure. And fourth, the subsidence history is calculated from the load history for different values of flexural rigidity. The difference between observed and calculated subsidence histories is minimized by altering the variation of strain rate through space and time. The calculated strain-rate variation is used to determine the temporal and spatial variation of heat flow into the sedimentary pile. The heat-flow variation constrains the temperature history of the sedimentary pile. In one dimension, the steady-state approximation yields:

$$T(z,t) = T_0(t) + Q(t) \int_0^z \frac{dz'}{k(z')} \quad (6)$$

where $T(z,t)$ is the temperature history, $T_0(t)$ is the surface temperature, $Q(t)$ is heat flow and $k(z)$ is the thermal conductivity of the compacting pile of sediment. When sedimentation is rapid and the strain rate is fast, the transient solution can be included^{36,40}.

ments and Quaternary fault slip data are modelled using inverse theory to obtain horizontal components of the instantaneous strain-rate tensor of the margin^{24,25}. Unfortunately, deformation has ceased at most continental margins, and indirect methods must be used to track margin growth, which is governed by the temporal and spatial variation of the strain rate tensor (see Box 1). This tensor determines three factors: (1) how the detailed shape of a margin evolves; (2) the degree of decompression melting; and (3) the variation of heat flow through time and space. Strain rate is the essence of the kinematic problem, which is solely concerned with the motion of particles. The more complete but difficult dynamic problem addresses how force acts upon materials to produce deformation²⁶.

We can determine the vertical component of the strain-rate tensor by inverting subsidence data from shallow sedimentary basins and margins where the stratigraphic record is known and where ancient water depths are reasonably well constrained^{27–30}. Existing one-dimensional and two-dimensional inverse algorithms assume that lithospheric stretching does not change with depth and do not include short-wavelength faulting. Future implementations will include depth dependency, which is probably important at highly stretched margins^{18,31}. In Fig. 4a, the results of inverting subsidence observations from more than 2,000 boreholes and field-measured sections demonstrate that strain rate varies by several orders of magnitude. At slow strain rates (10^{-17} – 10^{-16} s⁻¹), significant thermal diffu-

sion occurs as the base of the lithosphere is advected upwards. As a consequence, heat flow will vary little with time. At fast strain rates (10^{-15} – 10^{-14} s⁻¹), advection outpaces diffusion, and heat flow increases rapidly until stretching stops.

At margins, we expect the strain rate to increase towards the continent–ocean transition, because strain rates are $\sim 10^{-14}$ s⁻¹ at mid-oceanic ridges³². During the stretching phase, the increase in heat flow across the margin depends upon the strain rate. Once stretching ceases, the spatial variation of heat flow becomes more dependent upon the concentration of radiogenic heat components in the crust (Fig. 4b). If concentrations are low, post-rift heat flow will generally increase across the margin. If concentrations are high, post-rift heat flow will generally decrease across the margin. Thus, if we avoid complications resulting from the emplacement of hot melt, we expect heat-flow patterns to be predictable provided that the strain-rate history and the crustal composition are known. In the absence of borehole data, models are tested by comparing calculated and measured present-day heat flow (Fig. 4b). It is straightforward to use the spatial and temporal variation of heat flow to calculate the temperature history of the sedimentary pile (Fig. 5).

Hydrocarbon maturation

Once we understand how a margin has evolved, the temperature history of the sedimentary pile can be used to calculate the transforma-

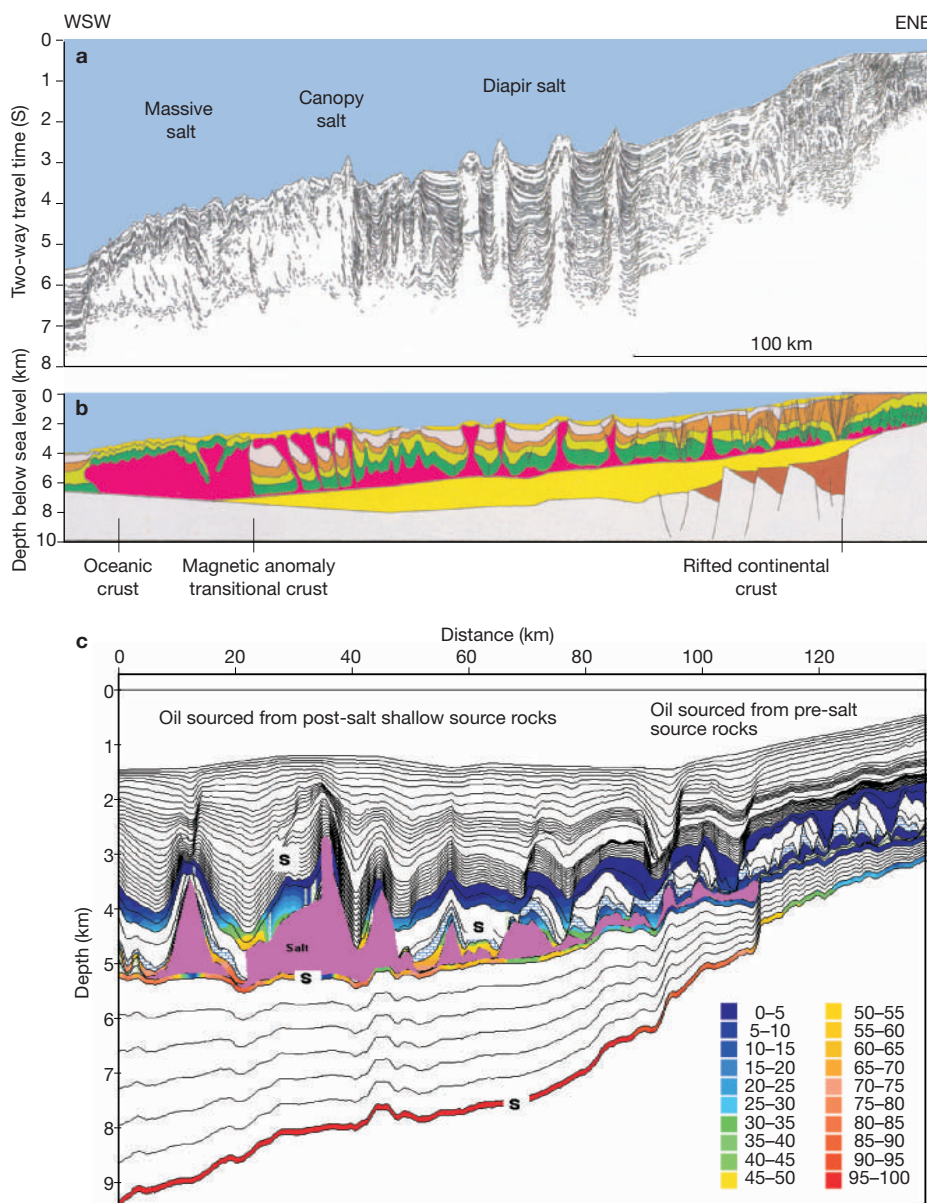


Figure 6 Structural and thermal maturation offshore of West Africa. **a**, Line drawing of a seismic reflection profile that traverses the West African continental margin from shoreline to abyssal plain (redrawn from ref. 34; see Fig. 2 for location). **b**, Simplified interpretation of the profile; yellow, white and green colours indicate strata that mostly consist of sandstones and shales; the pink layer indicates a deformed salt layer. Note the rapidly increasing water depth and anticlinal structures generated by deformation of salt. **c**, Portion of the profile that shows the calculated maturation of a set of source rocks at different depths. Warmth of colour indicates percentage of kerogen converted to hydrocarbons (blue, immature; yellow, mature; red, overmature). S, source rock.

tion of organic matter into oil and gas. These maturation calculations rely upon our understanding of the chemical kinetics of the many organic reactions that occur within solid organic matter (that is, kerogen) as it is transformed into oil and gas. We have a good quantitative understanding of the progress of these reactions, thanks to a combination of laboratory experiments and basin studies³³. Details of the maturation process are still poorly understood, but, from an exploration perspective, the key uncertainty remains the determination of source-rock temperatures in deeper waters away from borehole control. If there are different source rocks at different depths, changes in their temperature histories will affect the degree and timing of maturation, which can radically alter the dynamics of the hydrocarbon system.

A detailed description of the composition, porosity and permeability of strata combined with the spatial and temporal history of deformation is used to predict loci into which hydrocarbons have migrated. Once suitable traps have been identified, three-dimensional seismic imaging techniques can be used, under favourable cir-

cumstances, to locate accumulations (Box 2).

When deep-water exploration started on the West Africa Margin, the classic reservoir targets in water depths of less than 200 m were Cretaceous limestones and sandstones in rafted blocks. These reservoir rocks were charged by hydrocarbons generated by maturation of a source rock that lies deeply buried beneath a deformed salt layer (Fig. 6a, b)³⁴. One of the first deep-water boreholes in which hydrocarbons were encountered was within much shallower Cenozoic strata, which demonstrated that regional thermal gradients are much higher than had been expected (40–55 °C km⁻¹ rather than 25–35 °C km⁻¹). Thus the shallower reservoirs have been charged by hydrocarbons generated by significantly shallower source rocks.

A thermal maturation model of part of the West Africa Margin is shown in Fig. 6c. This model has been calibrated with maturation data from individual boreholes, and uses the predicted heat-flow history to calculate present-day maturation elsewhere. It clearly shows how present-day maturation of each source rock varies across the deepening margin. Understanding the spatial and temporal varia-

Box 2

Seeing is believing

Seismology is the principal method for imaging the solid Earth. It is of particular importance to the hydrocarbon industry, which uses acoustic energy generated by airguns to obtain high-resolution images of the Earth's subsurface down to a depth of ~10 km. It is now standard practice to acquire three-dimensional data that cover areas of ~10³ km². Seismic reflection data are the primary means for mapping the structure and composition of buried strata. Under certain circumstances, these data can also be used to detect the presence and composition of fluids trapped within the pore spaces of reservoir rocks. Subhorizontal reflections that cross-cut tilted strata are often seen on vertical slices cut through the image cube (Fig. 7a). These 'flat spots' can be generated by the change in acoustic impedance (that is, the product of density and velocity), which occurs at the contact between rocks containing two different fluids (for example, gas, oil or brine).

To determine the composition of both fluids, we must measure the change in acoustic impedance that occurs when sound waves are reflected from the 'flat spot' at different angles of incidence. The physical principles underlying this technique are well known. At any boundary, the reflection and transmission coefficients vary with the angle of incidence or offset. Zoeppritz⁴¹ formulated a set of equations that can be used to determine these coefficients as a function of offset and of elastic media properties (density, P-wave velocity and S-wave velocity). The Zoeppritz equations apply to the reflection of plane waves at a horizontal boundary between two half spaces and do not include complications caused by small-scale layering. They are nonlinear, and it is usually more convenient to use an approximation. Aki and Richards⁴² showed that if the change in elastic properties across the boundary is small, the Zoeppritz equations reduce to:

$$R(\theta) = \frac{1}{2} \left(\frac{\Delta \rho}{\rho} \right) \left\{ 1 - 4 \left(\frac{V_s^2}{V_p^2} \right) \sin^2 \theta \right\} + \frac{1}{2} \left(\frac{\Delta V_p}{V_p} \right) \sec^2 \theta - 4 \left(\frac{V_s^2}{V_p^2} \right) \left(\frac{\Delta V_s}{V_s} \right) \sin^2 \theta \quad (7)$$

where $R(\theta)$ is the reflectivity at angle θ , ΔV_p , V_p , V_s , ΔV_s , $\Delta \rho$, ρ and θ are the difference in P-wave velocity, the average P-wave velocity, the average S-wave velocity, the difference in S-wave velocity, the difference in density, the average density and the average of the

angles of incidence and of transmission. In the hydrocarbon industry, further simplifications are usually made, the most commonly used one being that of Shuey⁴³. An algorithm for calculating the variation of amplitude with offset for different rock properties is located at <http://www.crewes.org/Samples/ZoeppExp/ZoeppritzExplorer.html>. The acoustic properties of porous sedimentary rocks are estimated using empirical relationships⁴⁴.

At sea, reflected acoustic energy is recorded by hydrophones located along cables, which are up to 12 km long. Thus offset-dependent reflectivity is relatively easy to measure, provided all other sources of amplitude distortion are carefully removed. In other words, seismic data must first be processed to remove the effects of transmission loss, source and receiver response, spherical divergence and energy reverberation to isolate $R(\theta)$, which can then be compared with theoretical predictions. Figure 7b and c illustrates how the amplitude of a hydrocarbon-filled reservoir changes dramatically with offset. In this case, variation of reflectivity with offset is consistent with an oil–water contact.

Reflectivity analysis plays an important role in reducing exploration risk, but it is important to realize that the resultant 'direct hydrocarbon indicators' (DHIs) are often ambiguous⁴⁵. This ambiguity is a consequence of the trade-off between different acoustic properties (that is, many different combinations of fluid and rock properties can produce similar effects). A significant drawback is that small amounts of gas can have a dramatic effect on the variation of reflectivity with offset. An additional limitation is that deep exploration targets can occur beneath the DHI 'floor', where the acoustic discrimination of different fluid types is compromised by increased lithostatic pressure. Unfortunately, there are also examples of 'flat spots', which have the reflectivity characteristics of oil–water contacts but are caused by mineralogical phase changes⁴⁶. Nevertheless, offset-dependent reflectivity techniques have been spectacularly successful in deep-water exploration. On the West Africa Margin, BP has participated in drilling over 50 exploration wells between 1995 and 2003 with a commercial success rate of over 85%. This astonishing result is largely attributable to high-quality seismic imagery, which enabled oil pools to be identified with a high degree of certainty.

tion of this pattern in conjunction with the timing of salt-related deformation is crucial to ensure that suitable traps existed when hydrocarbons were expelled from the source rocks. The big question is what happens at the transition zone between thinned continental and bona fide oceanic crust. As we are dealing with either a 'cold' or 'warm' margin, it is likely that anomalous thermal gradients reflect the high levels of crustal heat-producing components (the conductivity of salt only plays a modifying role in focusing heat). Thus, present-day heat flow probably decreases towards the oceanic crust, and we expect that the source rocks shown in Fig. 6c will become steadily less mature as they occur further west. A key unknown is the temporal variation of heat flow, which might have been distributed to favour deep-water prospectivity. This uncertainty emphasizes the importance of a broader quantitative understanding of the margin's thermal and structural evolution.

The large thermal gradients encountered beneath the deep-water margin of West Africa are crucial to the success of shallow Cenozoic prospectivity. Many poorly explored margins, which have been written off because of mediocre results in shallow water, could also have zones of anomalously high heat flow. A campaign of low-cost shallow coring might be used to measure heat flow beneath the seabed and to

identify hydrocarbon seepage. If coring programmes were combined with the seismic imaging and modelling techniques described here, prospective margins with working source-rock systems could be identified.

Once a large discovery has been made, the production of the field is primarily an engineering problem, although in the past five years considerable use has been made of time-lapse acoustic imaging. Repeat three-dimensional surveys are carried out while the field is being used to produce oil, and, under favourable circumstances, it is possible to calibrate the amplitude response of subtle changes in hydrocarbon pressure and saturation. Thus the evolution of the field can be monitored and used to help design and fine-tune production.

Future directions

Deep-water exploration for hydrocarbons can be a high-risk strategy, but the conservative approach of drilling margins where there is a working hydrocarbon system has so far proved extremely effective. There have been spectacular successes offshore of the Gulf of Mexico and West Africa as well as in the Nile and Niger Delta areas (for example, Fig. 7). Crustal and lithospheric modelling based upon seismic imaging can help to address one important source of risk, namely, whether or not

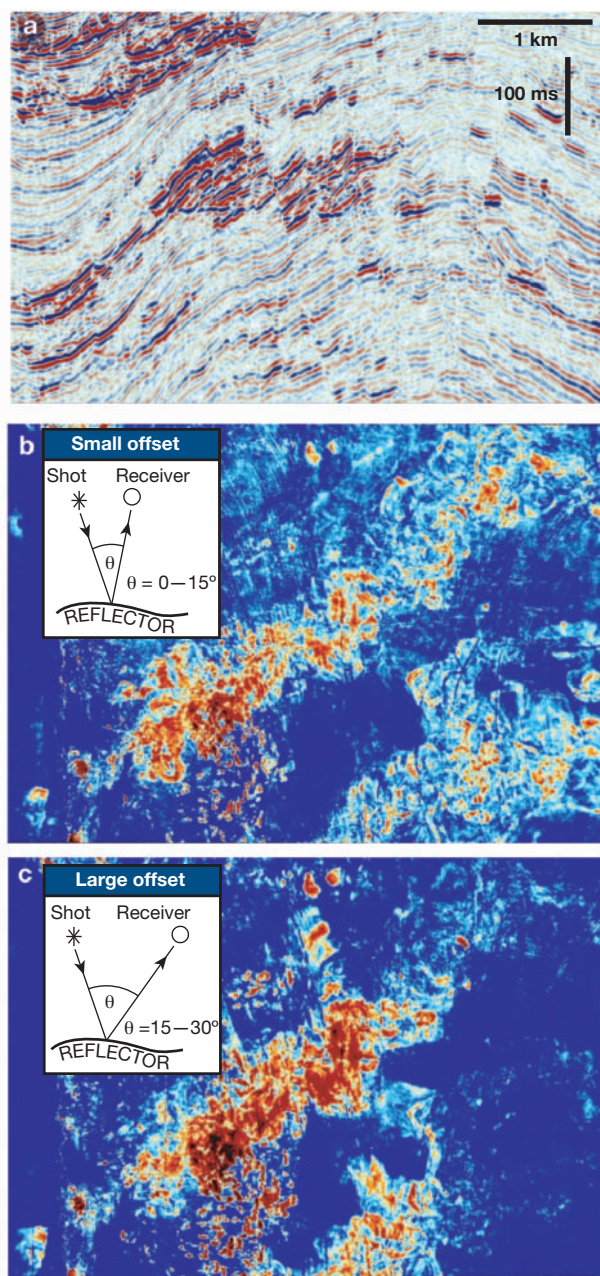
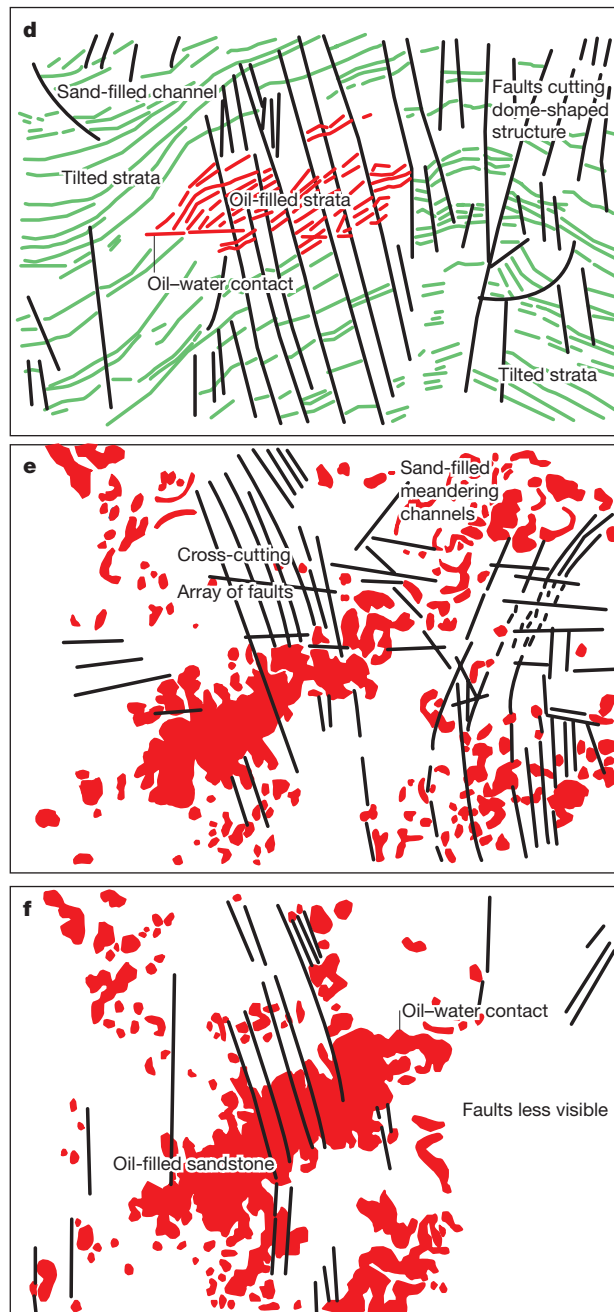


Figure 7 Three-dimensional imaging of an oil field. Set of images from a three-dimensional seismic survey acquired over a recently discovered hydrocarbon field that is located in 1,500 m of water offshore of West Africa. **a**, Vertical slice through an image cube on which folded and faulted strata form a large dome-shaped structure. Nearly half-way down the cross-section, a horizontal reflection can be seen, which cuts across stratal reflections and which is itself cut by a steeply dipping fault. This reflection is the oil–water contact. Tilted strata, which are located immediately above the oil–water contact and thus filled with oil, are much more reflective than those located below the contact. **b**, Horizontal slice cut through a three-dimensional image cube at the level of the oil–water contact (approximately 24 km²). This ‘amplitude extraction’ image was constructed using acoustic energy, which was reflected by small angles of incidence (0–15°; see inset sketch showing propagation of acoustic energy). Lithological contrasts are clearly seen: sets of yellow-orange-red (that is, high amplitude) meandering sand-filled submarine



channels constitute the hydrocarbon reservoir; the bulk of the continental slope is dominated by blue (that is, low amplitude) mudstones. Groups of intersecting ‘scratch marks’ are small faults that cut reservoir rocks. **c**, An image similar to **a** and **b** but constructed using acoustic energy which was reflected at larger angles of incidence (15–30°; see inset sketch). These ‘far-offset’ data show interstitial fluids. Orange-red colours (that is, high amplitudes) are oil-filled sandstones; yellow colours (that is, low amplitudes) are water-filled sandstones. Amplitudes of two visible channel systems are reduced, especially to the northeast where they are filled with water. The pattern of brightening and dimming coheres with the overall dome-shaped structure so that oil-filled sand located near the crest is brighter than brine-filled sand located on the dipping flanks. This ‘conformance to structure’ and the flat spot itself constitute the two principal direct hydrocarbon detection techniques. **d–f**, Sketches of all three seismic images that highlight the most significant features.

source rocks have reached thermal maturity at a suitable time with respect to trap formation. The hydrocarbon industry is now on the verge of exploring water depths that exceed 3 km, where the transitional zone between the continental and oceanic crust is encountered. Our ability to locate hydrocarbon accumulations in these water depths will largely depend upon understanding the thermal and structural evolution of these zones. A quantitative understanding will only emerge if ambitious and expensive seismic experiments are carried out at conjugate margin systems worldwide. These experiments will combine dense wide-angle and deep-reflection data and exploit both P-wave and S-wave acoustic energy. At the same time, there is a requirement for three-dimensional inverse algorithms that model the growth of margins.

Next summer, a group of scientists from the Universities of Southampton, Cambridge and Dublin in collaboration with BP plan to carry out an integrated and densely sampled seismic experiment across a young conjugate margin system in the Black Sea. The wide-angle and deep-reflection profiles will be combined with shallow seismic and borehole databases to yield highly resolved images that will illuminate the dynamic growth of a conjugate margin system.

Exploration of deep-water margins has just begun in earnest, and many margins are under-explored. Careful imaging and modelling at different scales will be crucial in the quest to identify the next generation of prolific margins. □

doi:10.1038/nature02133

1. C. Kuykendall, M. K. Hubbert and his Heirs: A Hubbert Peak Half-Biography at <http://www.oilcrisis.com/library/hubbeir.pdf> (2003).
2. The Association for the Study of Peak Oil & Gas <http://www.peakoil.net>
3. Parrish, J. T. *Palaeo-upwelling and the Distribution of Organic-Rich Rocks*. 199–205 (Spec. Publ. 26, Geol. Soc. Lond., 1987).
4. Milliman, J. D. & Syvitski, J. P. M. Geomorphic tectonic control of sediment discharge to the ocean: the importance of small mountainous rivers. *J. Geol.* **100**, 525–544 (1992).
5. Kuenen, P. H. & Migliorini, C. T. Turbidity currents as a cause of graded bedding. *J. Geol.* **58**, 91–127 (1950).
6. Douglas-Westwood & Infield Systems. The World Deep-Water Report IV 2003–2007 at <http://www.infield.com/Deepwater-Report2003-2007.html> (2003).
7. Steckler, M. S. & Watts, A. B. Subsidence of the Atlantic continental margin off New York. *Earth Planet. Sci. Lett.* **41**, 1–13 (1978).
8. McKenzie, D. P. Some remarks on the development of sedimentary basins. *Earth Planet. Sci. Lett.* **40**, 25–32 (1978).
9. Le Pichon, X. & Sibuet, J.-C. Passive margins: a model of formation. *J. Geophys. Res.* **86**, 3708–3720 (1981).
10. Loudon, K. E., Sibuet, J.-C. & Foucher, J.-P. Variations in heat flow across the Goban Spur and Galicia Bank continental margins. *J. Geophys. Res.* **96**, 16131–16150 (1991).
11. Loudon, K. E., Sibuet, J.-C. & Harmegnies, F. Variations in heatflow across the ocean-continent transition in the Iberian abyssal plain. *Earth Planet. Sci. Lett.* **151**, 233–254 (1997).
12. Foucher, J. P. et al. Heatflow in the Valentin Trough: geodynamic implications. *Tectonophysics* **203**, 77–97 (1992).
13. White, R. S. & McKenzie, D. Magmatism at rift zones: the generation of volcanic continental margins and flood basalts. *J. Geophys. Res.* **94**, 7685–7729 (1989).
14. Holbrook, W. S. et al. Mantle thermal structure and active upwelling during continental break-up in the North Atlantic. *Earth Planet. Sci. Lett.* **190**, 251–266 (2001).
15. Nottvedt, A. et al. *Dynamics of the Norwegian Continental margin*. (Spec. Pub., Geol. Soc. Lond., 2000).
16. Loudon, K. E. & Chian, D. The deep structure of non-volcanic rifted continental margins. *Phil. Trans. R. Soc. Lond.* **357**, 767–805 (1999).
17. Whitmarsh, R. B., Manatschal, G. & Minshull, T. A. Evolution of magma-poor continental margins from rifting to sea-floor spreading. *Nature* **413**, 150–155 (2001).

18. Minshull, T. A. The break-up of continents and the formation of new ocean basins. *Phil. Trans. R. Soc. Lond.* **360**, 2839–2852 (2002).
19. Bown, J. W. & White, R. S. in *Rifted Ocean-Continent Boundaries* (eds Banda, E., Torné, M. & Talwani, M.) (Kluwer, Dordrecht, 1995).
20. Chian, D. & Loudon, K. E. Crustal structure of the Labrador Sea conjugate margin and implications for the formation of non-volcanic continental margins. *J. Geophys. Res.* **100**, 24239–24253 (1995).
21. Holbrook, W. S. & Kelemen, P. B. Large igneous province on the US Atlantic margin and implications for magmatism during continental breakup. *Nature* **364**, 433–464 (1993).
22. Bauer, K. et al. Deep structure of the Namibia continental margin as derived from integrated geophysical studies. *J. Geophys. Res.* **105**, 25829–25853 (2000).
23. Watts, A. B. & Stewart, J. Gravity anomalies and segmentation of the continental margin offshore West Africa. *Earth Planet. Sci. Lett.* **156**, 239–252 (1998).
24. Haines, A. J. & Holt, W. E. A procedure for obtaining the complete horizontal motions within zones of distributed deformation from the inversion of strain rate data. *J. Geophys. Res.* **98**, 12057–12082 (1993).
25. Holt, W. E., Chamot-Rooke, N., Le Pichon, X., Haines, A. J., Shen-Tu, B. & Ren, J. The velocity field in Asia inferred from Quaternary fault slip rates and GPS observations. *J. Geophys. Res.* **105**, 19185–19210 (2000).
26. Bassi, G. Relative importance of strain rate and rheology for the mode of continental extension. *Geophys. J. Int.* **122**, 195–210 (1995).
27. Newman, R. & White, N. The dynamics of extensional sedimentary basins: constraints from subsidence inversion. *Phil. Trans. R. Soc. Lond.* **357**, 805–830 (1999).
28. White, N. An inverse method for determining lithospheric strain rate variation on geological timescales. *Earth Planet. Sci. Lett.* **122**, 351–371 (1994).
29. Bellingham, P. & White, N. A general inverse method for modelling extensional sedimentary basins. *Basin Res.* **12**, 219–226 (2000).
30. White, N. & Bellingham, P. A. Two-dimensional inverse model for extensional sedimentary basins: 1. Theory. *J. Geophys. Res.* **107**, doi10.1029/2001JB000173 (2002).
31. Davis, M. & Kusznir, N. Are buoyancy forces important during the formation of rifted margins? *Geophys. J. Int.* **149**, 524–533 (2002).
32. Kremer, C., Haines, J., Holt, W. E., Blewitt, G. & Laval, D. On the determination of a global strain rate model. *Earth Planets Space* **52**, 765–770 (2000).
33. Burnham, A. K. & Sweeney, J. J. A chemical kinetic model of vitrinite maturation and reflectance. *Geochim. Cosmochim. Acta* **53**, 2649–2657 (1989).
34. Marton, L. G., Tari, G. C. & Lehmann, C. T. in *Atlantic Rifts and Continental Margins* Geophys. Monogr. 115 (eds W. Mohriak & M. Talwani) (Am. Geophys. Union, 2000).
35. Wessel, P. & Smith, W. H. F. New version of the Generic Mapping Tools released. *Eos* **76**, 329 (1995).
36. Malvern, L. E. *Introduction to the Mechanics of a Continuous Medium* (Prentice-Hall, Old Tappan, NJ, 1969).
37. England, P. & McKenzie, D. A thin viscous sheet model for continental deformation. *Geophys. J. R. Astron. Soc.* **70**, 295–321 (1982).
38. Beavan, J. & Haines, J. Contemporary horizontal velocity and strain rate fields of the Pacific-Australian plate boundary zone through New Zealand. *J. Geophys. Res.* **106**, 741–770 (2001).
39. Lucaczeu, F. & Le Douarin, S. The blanketing effect of sediments in basins formed by extension: a numerical model. Application to the Gulf of Lion and Viking Graben. *Earth Planet. Sci. Lett.* **74**, 92–102 (1985).
40. Gallagher, K., Ramsdale, M., Lonergan, L. & Morrow, D. The role of thermal conductivity measurements in modelling thermal histories in sedimentary basins. *Mar. Petrol. Geol.* **14**, 201–214 (1997).
41. Zieppert, K. Erdbebenwellen VIII B, Über Reflexion und durchgang seismischer wellen durch unStetigkeitsflächen. *Göttinger Nachr.* **1**, 66–84 (1919).
42. Aki, K. I. & Richards, P. G. *Quantitative Seismology* (W.H. Freeman, New York, 1980).
43. Shuey, R. T. A simplification of the Zoeppritz equations. *Geophysics* **50**, 609–614 (1985).
44. Gassmann, F. Über die Elastizität poröser Medien. *Veierteljahr Naturforsch. Ges. Zurich* **96**, 1–23 (1951).
45. Rutherford, S. R. & Williams, R. H. Amplitude-versus-offset variations in gas sands. *Geophysics* **54**, 680–688 (1989).
46. Pegrum, R. M., Odegard, T., Bonde, K. & Hamann, N. E. Exploration in the Fylla area, SW Greenland. *Am. Assoc. Petrol. Geol.* (Regional Conference St Petersburg, 15–18 July, 2001).

Acknowledgements We thank K. Gallagher for peer review. S. Jones allowed us to show his unpublished basin animations (Fig. 5). A. Butler, R. Hardy, S. Jones, G. Kirby, D. Lyness, B. Lovell, M. Mayall, T. Minshull and J.-C. Sempere provided substantial help. We are grateful to BP, Sonangol and partners for permission to publish this paper and especially the seismic section and amplitude maps of Fig. 7. Some figures were prepared using GMT.

Biological activity in the deep subsurface and the origin of heavy oil

Ian M. Head, D. Martin Jones & Steve R. Larter

NRG petroleum group, School of Civil Engineering and Geosciences, University of Newcastle upon Tyne, Newcastle, NE1 7RU, UK (e-mail: steve_larter@hotmail.com)

At temperatures up to about 80 °C, petroleum in subsurface reservoirs is often biologically degraded, over geological timescales, by microorganisms that destroy hydrocarbons and other components to produce altered, denser 'heavy oils'. This temperature threshold for hydrocarbon biodegradation might represent the maximum temperature boundary for life in the deep nutrient-depleted Earth. Most of the world's oil was biodegraded under anaerobic conditions, with methane, a valuable commodity, often being a major by-product, which suggests alternative approaches to recovering the world's vast heavy oil resource that otherwise will remain largely unproduced.

Biodegraded oils dominate the world petroleum inventory¹, with the largest oil reserves being found, not in the Middle East, but on the flanks of foreland basins in the Americas (Box 1 Fig. 1). Biodegraded oils also represent a significant fraction of the petroleum in conventional oil reserves and will be common among future oil discoveries likely to be made in deep-water areas of the world (for example, the Atlantic margin basins of Africa, South America, Canada and the Gulf of Mexico). In these areas, low thermal gradients and shallow exploration targets combine to produce many petroleum reservoirs cooler than 80 °C, which is an ideal environment for microbial degradation of crude oil and natural gas².

The presence and role of microorganisms in deep subsurface sediments have been the subject of much interest recently (Box 2), despite the fact that geologists have had evidence of active microbial communities in petroleum reservoirs since the 1930s^{3,4}. Plate-tectonics, basin and petroleum-system formation are vital processes that affect deep subsurface biological systems. They control the formation and distribution of organic carbon, the delivery of nutrients and oxidants that can be used by microorganisms in the subsurface, and alter the physical environment that they inhabit. Plate tectonics, basin and petroleum-system formation are thus integral to the behaviour of the deep biosphere, and interactions between the biosphere and the geosphere in petroleum reservoirs provide a unique portal into the operation of life in Earth. Here, we summarize current knowledge of the composition of heavily biodegraded oils, factors that dictate their composition, and how the interplay of geological and biological phenomena influence the occurrence of biodegraded petroleum systems in the deep subsurface.

Effects of biodegradation on petroleum properties

Compositional changes, such as sequential and systematic removal of various hydrocarbons and other compounds, selective degradation of specific isomers within individual compound classes and the production of acidic compounds, seen in shallow petroleum reservoirs, are similar to those seen in surface oil seeps and laboratory biodegradation experiments. These changes are distinct from alterations caused by physical processes, such as water washing or phase fractionation, and suggest biological alteration of oils *in situ*^{2,5-14}.

The effects of biodegradation on the composition and physical properties of crude oil and natural gases are well known. Oxidation of oil (C₆₊ components) during biodegradation leads to a decrease in saturated hydrocarbon content (and to a smaller decrease in aromatic hydrocarbon content) and API gravity, a measure that correlates with economic value, whereas oil density, sulphur content, acidity, viscosity and metal content increase^{2,8,9,11,13,15,16} (Fig. 1), which negatively affects oil production (by reducing well-flow rates) and refining operations, and reduces oil value.

Hydrocarbons are preferentially destroyed during biodegradation, but sulphur-, oxygen- and nitrogen-containing compounds can also be degraded¹⁷⁻¹⁹. New compounds such as acyclic and cyclic, saturated and aromatic carboxylic acids and phenols are produced from hydrocarbons^{11,12,20}, and a complex variety of acidic non-hydrocarbons is generated from the aromatic heterocycles found in oil²¹. The heterocyclic acids with multiple, different heteroatoms are the main cause of corrosion problems during processing of heavy, degraded oils²².

Light hydrocarbons and gases (C₂–C₅ hydrocarbons) can also be biodegraded²³⁻²⁵; propane is degraded most rapidly, and *n*-butane is more susceptible to degradation than iso-butane (Fig. 1). Biodegradation of C₂–C₅ hydrocarbons increases the methane content, and this, with the reduction in oil alkane content during biodegradation, reduces the ability of the oil to dissolve gas, reducing the gas:oil ratio of any trapped oil leg (the oil-saturated part of the reservoir)¹³ (Fig. 1). This may result in production of a gas cap enriched in methane following removal of wet gas components (C₂–C₅ alkanes) and probably also by direct production of methane during biodegradation. Large dry gas caps associated with residual biodegraded oil are common^{24,26} (S. L. and Di Primio, unpublished data), and probably result directly from oil biodegradation. There is little direct evidence of methane oxidation during petroleum biodegradation *in situ*, but we recently recovered archaeal 16S rRNA gene sequences characteristic of anaerobic methane oxidizers from biodegraded oil reservoirs, suggesting that anaerobic methane oxidation occurs at some point during the subsurface degradation process.

All oils are mixtures

The composition of oil in reservoirs is the result of the processes occurring during petroleum generation, migration, trapping and subsequent alteration. Petroleum gener-

Box 1

The mothers of all petroleum systems

The world's oil reserves are dominated by biodegraded heavy and super-heavy oils in the super-giant tar sands common in shallow reservoirs on the flanks of foreland basins in North and South America and elsewhere.

Oils are classified for economic value according to API gravity, based on a surface measurement of the specific gravity of degassed oil¹⁶. Heavy oils have API gravities of 20 or less, super-heavy oils have API gravities of 10 or less, and a typical light marine non-biodegraded oil has an API gravity around 36–38 API. Tar sands are sandstones saturated with heavy or super-heavy oil: the oils in the Canadian and Venezuelan tar sands have API gravities of 6–12. The vast majority of heavy oils result from microbial alteration of oils in the reservoir^{1,15,16,37,84,85}, with over 50% of the Earth's oil inventory occurring as biodegraded oils in heavy oil and tar sand accumulations¹. The largest single accumulations are the supergiant deposits of tar sands trapped on the flanks of the Alberta (Canada) and Eastern Venezuelan (Venezuela) foreland basins^{1,84}. The Orinoco heavy oil belt in Venezuela has the largest known petroleum accumulation in the world (1,200 billion barrels — one barrel is 0.159 m³), with the Athabasca accumulation in Canada in second place (nearly 900 billion barrels). These accumulations dwarf the largest light oil fields such as Ghawar (Saudi Arabia) and Burghan (Kuwait), which contain a measly 190 billion barrels each¹.

Foreland basins are extensive (up to thousands of kilometres long, hundreds of kilometres wide), elongated basins, which are formed as a result of isostatic down warping of the crust, owing to the tectonic load imposed during a mountain-building (orogenic) event. When source rocks are present in the older platform sequences or in the foreland basins sediments themselves, these

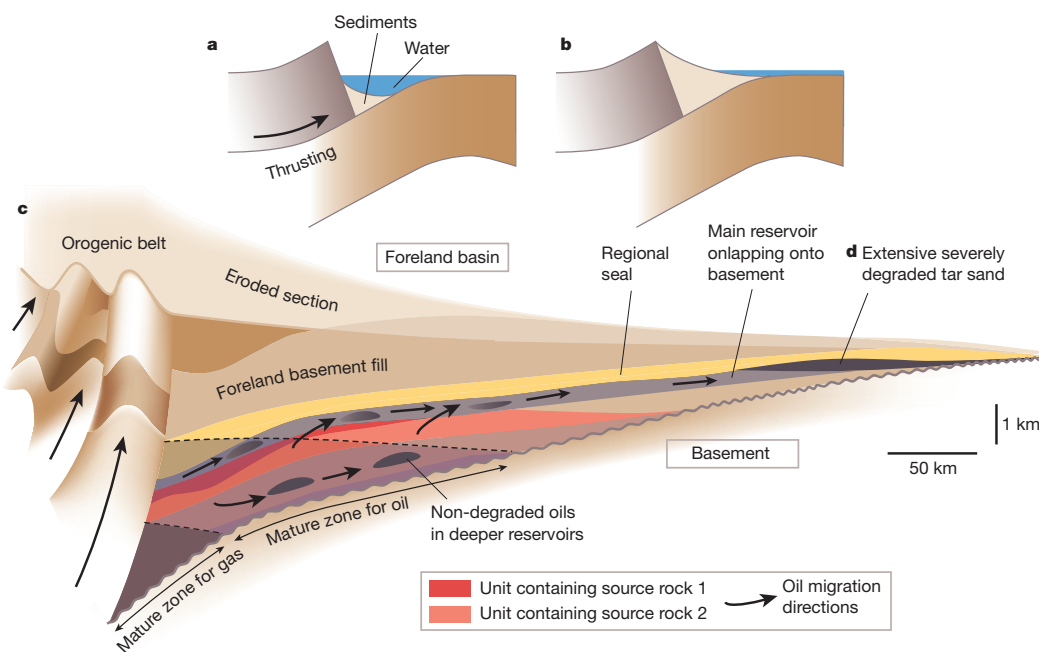
basins develop enormous large elongated oil source kitchens, which produce vast volumes of oil. This oil migrates laterally up the sloping strata from the source rocks for up to several hundred kilometres to structural and stratigraphic traps at the margins of the basin^{84–87}. Viable petroleum traps occur throughout foreland basins, and light oils may be trapped along the migration fairway, in reservoirs found deeper in the basin. At the flanks of the basins, reservoir sediments (typically sandstones) lay onto older basement rocks and produce reservoirs that are shallow, cool and may have local active meteoric water circulation⁵⁹, conditions ideal for biological activity. The formation of foreland basin petroleum systems feeds the deep biosphere at the basin flanks, where stratigraphic traps may accumulate vast volumes of biodegraded heavy oil trapped significantly by immobility owing to the high density and viscosity resulting from severe biodegradation (Box 1 Fig. 1).

In the case of the Western Canada Basin the Lower Cretaceous Mannville Group sandstones onlap onto subcropping Palaeozoic carbonates, which together contain in excess of 1,700 billion barrels of severely biodegraded heavy oils of 8–12 API in the Athabasca, Cold Lake, Wabasca and Peace River accumulations, representing over 97% of the Western Canada Basin oil reserves⁸⁵. More than 36% of Canadian oil production is currently derived from this resource⁵⁹ either by mining the near-surface deposits or by heating the oil by steam injection to reduce the viscosity and produce the oil through boreholes^{59,88}.

The source of the oils is controversial, although it is agreed they represent biodegraded mature marine oils charged from pre-Cretaceous source rocks during the development of the Rocky Mountain Fold and Thrust Belt (Laramide orogeny) in Late

Cretaceous–Early Tertiary time^{85,86,89,90}.

The Eastern Venezuelan Basin contains the 460 km by 40 km wide Orinoco Tar Belt, containing 1,200 billion barrels of severely degraded oil found predominantly in stratigraphic traps formed by onlap of Miocene sandstones onto the Guayana shield basement. Reservoirs were charged in the Neogene from Upper Cretaceous marine source rocks from up to 200 km away in the foreland basin depocentre created by formation of the Interior range⁸⁷. These vast foreland basin petroleum generation, accumulation and alteration engines show graphically the dynamic interplay of tectonics, petroleum geology and active biosphere.



Box 1 Figure 1 Foreland basin petroleum systems are major heavy oil provinces. An idealized foreland basin petroleum system with source rocks in the platform sediments, a regionally extensive reservoir sandstone sequence and a regionally extensive seal. Mature oil source rocks in the basin depocentre charge the flank traps where the oils biodegrade and tar sands accumulate. **a, b**, The foreland basin forms during the period of active deformation

and subsequently buries the source rocks present in the pre-foreland basin passive margin sequences and any source rocks formed in the foreland basin itself into the oil window (temperatures of ~100–150 °C). **c**, Oil migrates to the shallow cool basin flanks over distances of several hundred kilometres. Biodegradation of the oil occurs, increasing viscosities and helping to trap the degraded oil.

ation is the rate-determining step in charging reservoirs with new oil²⁷, and source rocks typically charge petroleum traps for timescales of the order of a few million to tens of millions of years. During this period, the petroleum expelled from the source(s) and subsequently trapped shows a progressive change in composition. This change results from alterations in the composition of the oil from the source rock(s) over time, temporal differences in migration-related phase fractionation effects and mixing of oils in the trap. Mixing of fresh and degraded oils dominates the composition and physical properties of biodegraded oils^{24,28–30}, rendering geochemical schemes that rank the level of biodegradation of oil unachievable. Nevertheless several utilitarian schemes have been proposed that have had wide and successful application in petroleum exploration^{2,8,13,31–33}, with the schemes of Volkman *et al.* (ref. 8) and Peters and Moldowan (ref. 32) being the standard descriptions of the extent of biodegradation of produced oil. These schemes are based heavily on alteration of biomarker compounds; however, much of the commercially significant changes in oil properties occur before significant alteration of cyclic biomarker hydrocarbons takes place. The more pragmatic scale of Wenger *et al.* (ref. 13) defines five broad levels (from slight to severe) of biodegradation and is used as the framework for Fig. 1.

Quantitative evaluation of oil biodegradation

All biodegraded oils are mixtures; therefore, estimating the fraction of oil lost through degradation is not simple. The most obvious biodegradation-resistant tracers of original oil mass, such as oil vanadium or nickel content, vary in concentration by orders of magnitude as oils are generated through the oil-generation window (the temperature range over which oil is generated and expelled from source rocks). Thus variations in oil–metal content through biodegradation may not accurately indicate oil mass lost during degradation when maturity of the oil charge varies. In an attempt to link qualitative indicators of oil biodegradation (for example, the Peters and Moldowan scale; Fig. 1) with quantitative estimates of the mass of oil destroyed, data on bulk oil composition and variations in the concentration of degradation-resistant compounds in oil (where little variation in the maturity of the charged oil has occurred) have been analysed. These data, used with oil-charging biodegradation models²⁹, suggest that oils exhibiting heavy levels of biodegradation (Fig. 1) have typically lost up to 50% of their mass of C₆₊ components. Beyond this level of degradation, loss of oil mass from marine oils is less significant (only a further ~10–20% of original oil mass is lost), and it seems likely that structural rearrangements dominate subsequent changes in oil composition. The extent to which degradation is resistant to microbial necromass addition to oil during biodegradation is unclear, but it seems likely that addition of microbial necromass occurs at some level.

Controls on biodegradation

Most petroleum (oil and gas) is trapped in reservoirs in porous and permeable sediments such as sandstones and limestones. Typically, in the oil leg, oil, as the continuous fluid phase, accounts for ~80% of the pore space, with discontinuous water filling the rest. Below the oil leg in the water-saturated part of the reservoir (the water leg), 100% of the pore space is saturated with waters of variable salinity (Fig. 2). Free water is essential for biological activity, and biodegradation of oil could, in principle, occur throughout a reservoir. Biodegradation could also occur during early reservoir filling, when many oil–water contacts exist²⁷. Significantly, however, degradation-related compositional gradients are seen in many oilfields (Fig. 2), with the most-degraded oil near oil–water contacts²⁹, or highly degraded residual oil²⁴. This suggests that the site of most biodegradation is at the base of the oil column. The oil–water contact provides conditions that are the most conducive to microbial activity. Transport of hydrocarbons from the oil column will provide a plentiful supply of electron donors, whereas inorganic nutrients required for microbial growth

can, in principle, be transported by water flow or diffusion in the water column to the biosphere at the oil–water contact. This notion is supported by our observation that 16S rRNA genes from bacteria and archaea are more readily detected in sediment samples from the vicinity of the oil–water contact.

Large-scale tectonic processes

Biodegraded oils are found at depths of up to about ~4 km (ref. 2), with most biodegraded reservoirs at up to 2.5 km below the sediment surface. Deep subsurface petroleum reservoirs are characterized by high temperatures, with the temperature increasing ~2–3 °C per 100 m depth. The probability of observing oil biodegraded to a given extent has been noted empirically to increase with decreasing reservoir temperatures below about 80 °C (refs 2, 15, 34, 35), with the sensitivity of degradation to temperature seemingly independent of oil type³⁴. Thus the probability of finding heavily degraded oils (level 5 on the Peters and Moldowan biodegradation scale; Fig. 1) is close to 0 for reservoirs near 80 °C, whereas for reservoirs near 50 °C, 70% of reservoirs might typically contain heavily degraded oil³⁴. Thus significant biodegradation decreases with increasing reservoir temperature and essentially ceases around 80 °C. Biodegraded oils are only rarely found in currently subsiding reservoirs above 80 °C, which suggests that most biodegraded oils in reservoirs in these basins at lower temperatures have been charged and degraded recently, close to their present depths of burial²⁹.

Not all low-temperature reservoirs contain degraded petroleum. This may be because either they have been recently charged with fresh oil or they have been uplifted from deeper, hotter subsurface regions. Wilhelms *et al.* proposed a ‘palaeopasteurization’ model in which these petroleum reservoirs were pasteurized at 80–90 °C during deep burial³⁵, inactivating any hydrocarbon-degrading microorganisms present, before the main oil charge and subsequent uplift of the reservoir to cooler conditions (Fig. 3). This implies an important role for large-scale geological processes in shaping the extent of the deep subsurface biosphere. An important implication of this model is that following uplift, recharge of fluids and/or microorganisms from the surface and migration of microorganisms into pasteurized deep subsurface petroleum reservoirs are insignificant. This also implies that in biodegraded/biodegrading petroleum reservoirs in currently subsiding basins cooler than 80 °C, microorganisms were already present during burial, having survived and evolved since reservoir deposition. Even in coarse-grained sediments deposited under oxygenated surface conditions, oxygen is rapidly consumed as a result of aerobic organic matter mineralization. Most shallow subsurface sediments therefore become anoxic with depth, and anaerobic microbial communities dominate most shallow subsurface sediments. The consumption of labile organic matter leads to sediments that are dominated by residual recalcitrant organic carbon harbouring an anaerobic microbial community. During burial of such sediments and formation of potential reservoir rocks, acetate may be produced biogenically from more recalcitrant organic matter with increasing depth and temperature³⁶. This may sustain an indigenous heterotrophic anaerobic microbial community until oil arrives in the reservoir. We conjecture that the base of oil biodegradation effectively marks the base of heterotrophic microbial activity in the Earth’s crust.

Sustaining petroleum degradation: oxidant sources

Aerobic biodegradation of hydrocarbons dominated thinking about the mechanisms of subsurface oil biodegradation for many years^{15,16,37}, despite conservative estimates of the water volumes needed to transport enough oxygen presenting overwhelming problems geologically in most petroleum reservoirs³⁸. In our experience, deep biodegraded oilfields in marine basins often contain saline water, indicating minimal flushing of deep aquifers with fresh water, and in the North Sea, reservoir water salinities in the largest biodegraded petroleum accumulation in the basin, the Troll field, are equal to or

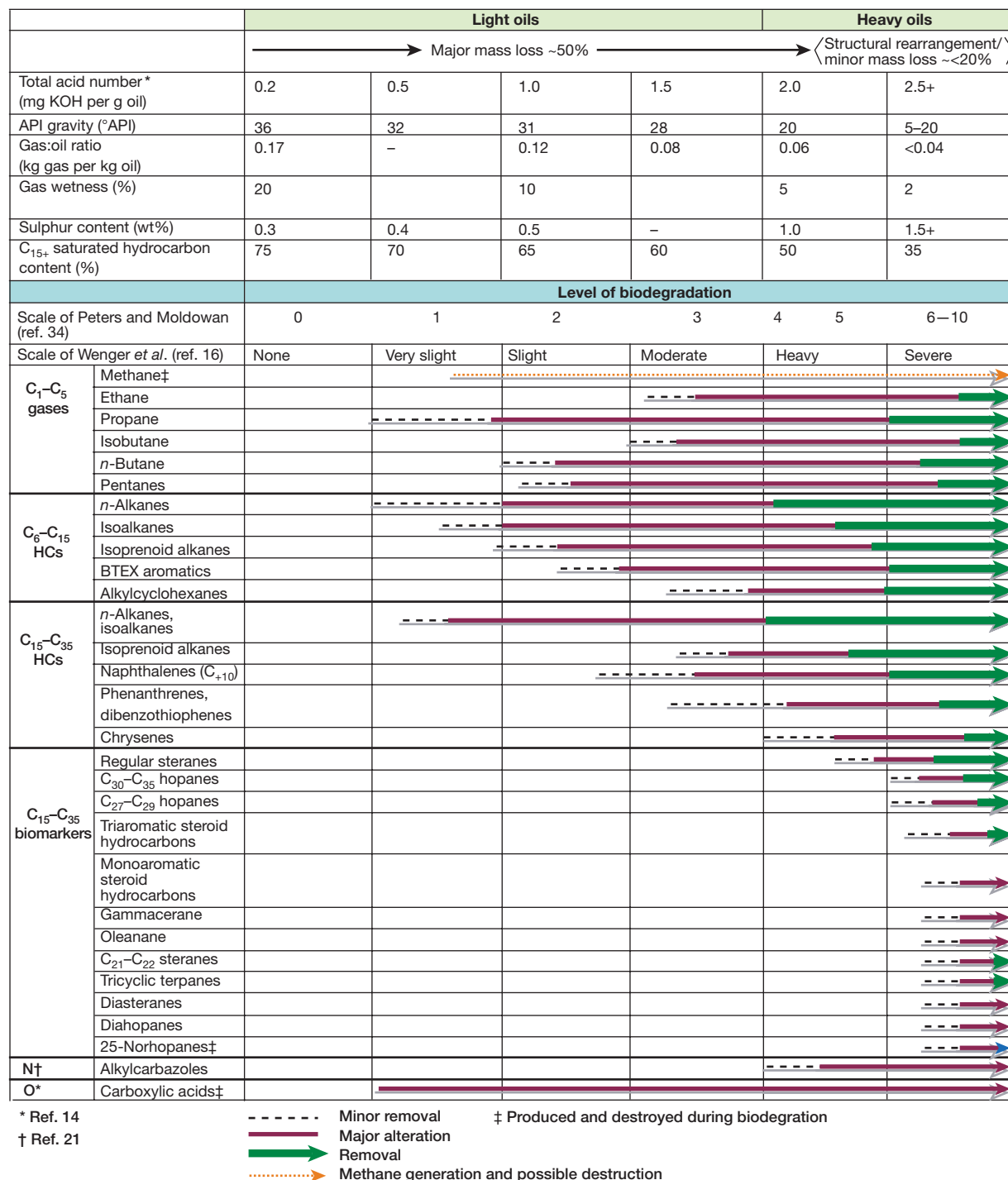


Figure 1 Schematic diagram of physical and chemical changes occurring during crude oil and natural gas biodegradation. The general sequence of removal of saturated hydrocarbon types during biodegradation is *n*-alkanes, alkylcyclohexanes, acyclic isoprenoid alkanes, bicyclic alkanes–steranes–hopanes⁷, with some production of new hydrocarbons, such as 17 α ,25-norhopanes from demethylation of hopanes at advanced levels of degradation^{32,33}. Similarly, for aromatic hydrocarbons, alkylbenzenes are removed before diaromatic and triaromatic hydrocarbons⁸, with aromatic steroid hydrocarbons being resistant until very severe levels of biodegradation are achieved³². With both alkanes and aromatic hydrocarbons, stereochemistry and the position of alkyl substituents confound such simplistic scenarios, and there is frequently much overlap and synchronous removal

of the various compound types. Removal of gas chromatography (GC)-resolvable components during degradation contributes to the production of the ‘humps’ of GC-unresolvable hydrocarbon mixtures that are typically found in degraded oils⁹¹. However, even when biodegradation levels are slight, mass balance calculations suggest that non-GC-resolvable components are also degraded. The indicative compositional and physical property changes with biodegradation assume that it starts with a light, low-sulphur undegraded oil. The lower diagram of compositional change is modified from Wenger *et al.* (ref. 13). Pervasive oil mixing and synchronous removal of several compound types result in the degradation scales being indicative rather than being a precise consistent form.

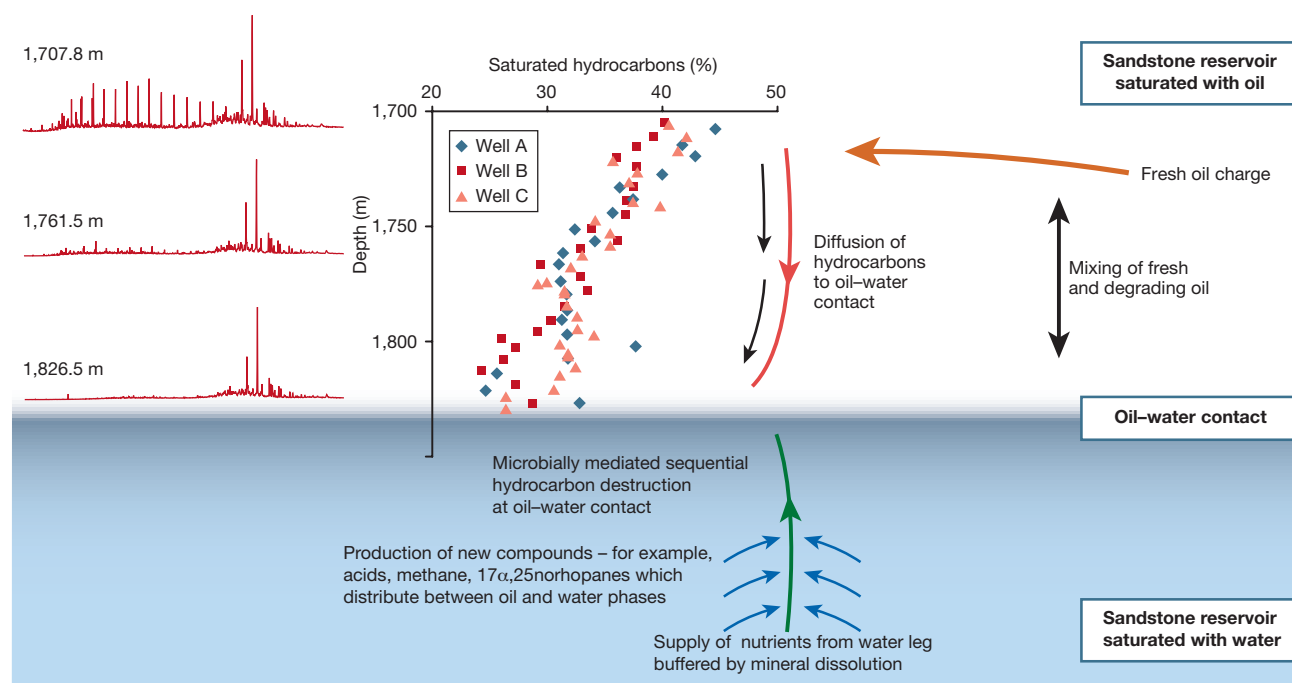


Figure 2 Saturated hydrocarbon contents and gas chromatograms of petroleum extracted from reservoir cores show a progressive increase in biodegradation in three wells from a Chinese oilfield¹⁹. Hydrocarbons diffuse towards the oil–water contact, where they are degraded by microorganisms living near the oil–water contact using nutrients derived from the water-saturated zone below the oil column.

Nutrients such as phosphorus are probably buffered by mineral dissolution reactions⁷². Fresh oil is charged to the reservoir at the same time that degradation occurs. Compositional gradients reflect this complex charge and degradation scenario.

greater than sea water³⁹. Recent microbiological evidence from our group and others suggests, even when reservoirs are shallow (< 500 m) and contain fresh water, that the microbial flora is anaerobic. There is strong evidence for the widespread occurrence of obligate anaerobes in subsurface petroleum systems^{40–42} (Box 3).

Flushing of meteoric water in itself does not indicate that highly reactive oxygen survives transport to deep reservoirs, past reactive organic matter and minerals, because even small concentrations of organic compounds can swiftly remove oxygen from an aquifer. It is therefore almost certain that oil biodegradation in deep subsurface petroleum reservoirs proceeds through anaerobic microbial metabolism rather than through aerobic mechanisms. This was proposed many years ago^{43,44}, but was not widely accepted in the face of overwhelming evidence that hydrocarbon degradation only occurred rapidly under oxic conditions. The first bacteria isolated from oil-field waters were anaerobes³, and the widespread occurrence of anaerobic bacteria and archaea in subsurface petroleum systems is supported by many more recent studies^{41,42,44–46} (Box 3). In addition, viable anaerobic hydrocarbon degradation processes have recently been established for both saturated and aromatic hydrocarbons^{47–49} (Fig. 4), and metabolites characteristic of anaerobic hydrocarbon degradation such as reduced 2-naphthoic acids⁵⁰ have been identified in most biodegraded oil reservoirs that we have examined, including the Canadian tar sands⁵¹. Although anaerobic bacteria have been widely reported in petroleum reservoirs, none of these are anaerobic hydrocarbon degraders, and only one thermophilic anaerobic hydrocarbon degrader potentially capable of living in the deepest biodegraded reservoirs has been identified⁵². It does seem certain, however, that hydrocarbon degraders inhabit deep reservoirs. The subsurface petroleum reservoir biosphere awaits the explorers of the twenty-first century.

Slow anaerobic processes dominate the deep subsurface, with subsurface hydrocarbon degradation probably linked to iron reduction and methanogenesis^{10,47,49,53} or, in petroleum reservoirs where

free sulphate is abundant, to microbial sulphate reduction⁵⁴. The high hydrogen sulphide concentrations produced may ultimately stop biodegradation¹⁴. Methanogenesis, an exclusively anaerobic process, appears to be commonly associated with biodegraded petroleum reservoirs that are free of abundant sulphate, and isotopically lighter methane is frequently found mixed with thermogenic methane^{25,26,30,55–57}. Methanogenesis is thought to have contributed to the gas associated with the Canadian tar sand deposits^{58,59}, and similar processes have been proposed for biogenic methane generation in coals^{55,57}.

The carbon isotopic composition $\delta^{13}\text{C}$ of methane associated with subsurface oil biodegradation conditions frequently seems to be in the range of -45‰ to -55‰ , but the carbon dioxide associated with biodegraded oils shows a wide range and is, in contrast, often isotopically heavy ($\delta^{13}\text{C}$ up to $+15\text{‰}$), which is indicative of almost complete closed-system reduction of carbon dioxide to methane^{25,26,30} (Fig 4).

Methanogens are probably indigenous members of petroleum reservoir microflora^{45,60,61}. The methanogens described so far are those that reduce carbon dioxide to methane, although there are a few reports of acetoclastic methanogens from petroleum reservoirs⁶². Radiotracer experiments indicate that carbon dioxide reduction to methane is more prevalent than acetoclastic methanogenesis⁶³, and the high pressures in petroleum reservoirs would be expected to favour net-volume-reducing reactions such as methanogenesis from carbon dioxide reduction rather than from acetate. Methanogenesis is a likely fate for most carbon dioxide produced during oil biodegradation in the absence of abundant sulphate, as little carbonate cementation is seen in many biodegraded oilfields despite large fractions of the petroleum carbon being oxidized. Carbon dioxide contents of biodegraded petroleum reservoirs are generally well below 10% by volume of the produced gas phase and are often not much higher than nearby non-degraded oilfields, which typically contain only a few mol % of carbon dioxide. Although hydrogen for initial

methanogenesis is probably derived from water⁴⁷, secondary reduction of the remaining carbon dioxide requires an additional source of hydrogen. This may be supplied from depth by mineral hydrolysis⁶⁴, maturation of organic matter or even from the aromatization of alicyclic or naphthoaromatic compounds present in the oil. The petroleum biosphere might have many of the characteristics of other hydrogenotrophic subsurface biota⁶⁵. We speculate that methanogenesis through carbon dioxide reduction may be a dominant terminal process in petroleum biodegradation in the subsurface⁴² (Fig. 4) and is probably active today in the giant Troll field in the North Sea²⁶.

Water itself is a likely reactant for the production of methane, carbon dioxide and free energy during hydrocarbon degradation in deep anoxic settings. Thermodynamic and microbiological evidence support this^{47,53,66}, and water is of course geologically the most plausible universal reactant in petroleum reservoirs. Oil plus water producing life — and a little gas (Fig. 4).

Sustaining petroleum degradation: inorganic nutrients

Although transport of oxygen to fuel hydrocarbon degradation in petroleum reservoirs seems unlikely, where water movement occurs it might enhance mineral dissolution and release nutrients such as phosphate, promoting microbial activity. Hydrocarbon degradation in surface environments is often limited by nitrogen or phosphorus, but nitrogen is unlikely to be limiting in petroleum reservoirs, because nitrogen in petroleum systems is as abundant as ammonium ions in water, and as dinitrogen gas and heterocyclic aromatic nitrogen compounds in petroleum^{19,42,67,68}. Aromatic petroleum nitrogen compounds are degraded in heavily degraded oils^{18,19}, but it is likely that ammonium ions buffered by reservoir minerals⁶⁷ are the primary nitrogen source for the hydrocarbon-degrading biosphere in petroleum systems⁶⁸, with phosphorus much more likely to limit oil biodegradation in petroleum reservoirs. In pivotal studies, Bennett and co-workers^{69–72} have shown a close relationship between the geomicrobiology of a petroleum-contaminated aquifer, mineral alteration and groundwater chemistry. On a large scale, biological activity perturbs general groundwater chemistry, and therefore mineral water equilibria, and on a small scale, attached organisms locally perturb mineral water equilibria, releasing limiting nutrients. In an oil-contaminated aquifer, feldspars weather exclusively near attached microorganisms in the anoxic region of the contaminant plume; indigenous bacteria colonize feldspars that contain potassium or

trace phosphorus as apatite inclusions. Most of the phosphorus in sediments from clastic petroleum reservoirs is in feldspars⁷⁰, and feldspar dissolution in some oil reservoirs (for example, the Gullfaks field in the North Sea) may be related to biodegradation of the associated oils⁷³. Indeed we think that supply of limiting nutrients (probably phosphorus) from mineral dissolution in reservoirs or encasing shales in many instances may be the rate-limiting step in subsurface petroleum biodegradation.

Rates of oil biodegradation in petroleum reservoirs

It has been suggested that significant oil biodegradation occurs naturally in a petroleum reservoir on a human timescale⁷⁴, and when sea water is injected as part of a production strategy, sulphate reduction may be stimulated⁴⁶. Under subsurface conditions, however, biodegradation is usually a slow process. We have assessed biodegradation rates using oilfield charge models and observed oil compositions, diffusion-controlled oil column compositional gradients and by examining mixed degraded/non-degraded oils in oilfields and determining the net degradation rate in terms of the oil charge rate²⁹. These three quite different approaches all estimate first-order biodegradation rate constants of between 10^{-6} to 10^{-7} yr⁻¹ for hydrocarbons in reservoirs near 60–70 °C, if it is assumed that biodegradation takes place in the lowest 2% of the oil column. Considering global levels of biodegradation in oil reservoirs, we find that hydrocarbon destruction fluxes at the base of oil columns are of the order of 10^{-4} kg hydrocarbons m⁻² yr⁻¹ for reservoirs at 40–70 °C.

Our models suggest that biodegradation takes around 1–2 Myr to compositionally perturb an entire oil column of 100 m for lightly degraded oil reservoirs, and where we see continuous compositional gradients in oil columns related to biodegradation, degradation must have been occurring episodically for many millions of years (Fig. 2). It seems that to remove the *n*-alkanes from oil, around 5–15 Myr is needed, depending on the oil properties, and we estimate the Canadian tar sands could have been degraded to their current state in as little as ~35 Myr. The timescales of oilfield degradation and filling are thus very similar, and the degree of biodegradation and the physical properties of oil will be substantially controlled by oilfield charge history and mixing^{29,75}.

The oil biodegradation rate is not limited by electron-donor supply (that is, hydrocarbons) but by supply of nutrients or oxidants. This suggests that supply of nutrients or electron acceptors to the site

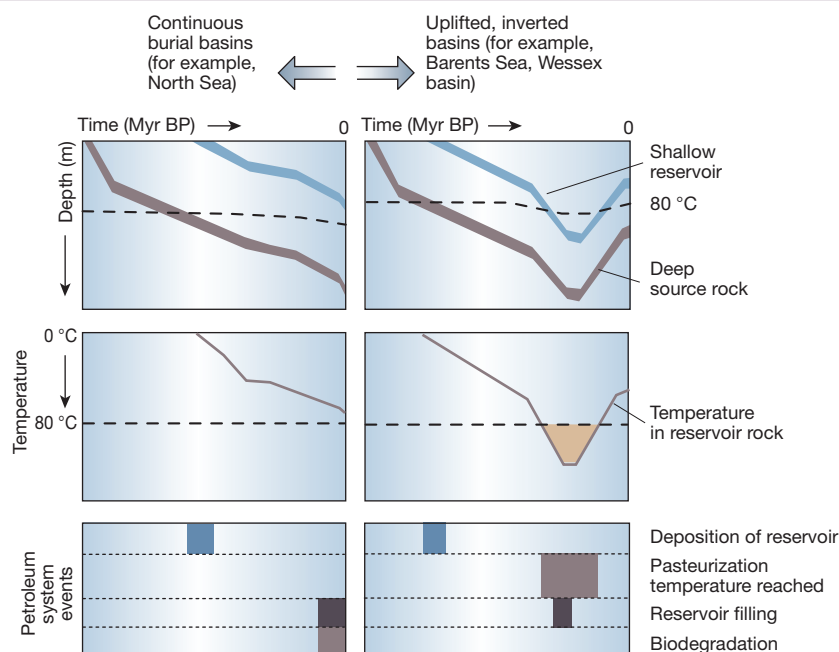


Figure 3 The palaeopasteurization model³⁵ of Wilhelms *et al.* compares continuously subsiding (for example, Viking Graben, North Sea) and uplifted sedimentary basins (for example, Barents Sea or Wessex Basin) and shows schematic burial history (top), reservoir temperature history (middle) and petroleum system (lower) events. This illustrates key differences with respect to biodegradation in petroleum reservoirs. Biodegradation of petroleum occurs only in sedimentary units that have not been exposed to temperatures exceeding 80 °C (palaeopasteurization) before oil charging. We speculate that this 'pasteurization temperature' may represent the effective maximum temperature boundary for most life in Earth struggling to survive over geological timescales in hot oligotrophic sediments.

Box 2

The deep biosphere

Although bacteria putatively associated with deep subsurface environments were characterized in the first part of the twentieth century, it is only recently that the existence of a deep subsurface biosphere has been widely accepted in the scientific community. Scepticism about the deep subsurface biosphere stemmed largely from concerns about the provenance of organisms detected in samples taken from the subsurface, but with improved sampling techniques and methods to corroborate the authenticity of organisms recovered from deep within the Earth's crust many of these concerns have been dispelled.

Soils and surface sediments can harbour billions of prokaryotic cells per cubic centimetre, but their abundance drops off exponentially with increasing depth⁸². Nevertheless, some deep subsurface sediments still hold around 10^5 to 10^6 cells cm^{-3} , and where carbon and energy sources are abundant, numbers may increase at depth. Evidence suggests that a deep crustal biosphere beneath both land and sea has reached approximately 3 km below the Earth's surface, with oil biodegradation suggesting that this can be extended to at least 4 km. Although the number of cells per unit volume of deep subsurface sediments is relatively small, the vast extent of the sediments (estimated at $5 \times 10^{25} \text{ cm}^3$) means that total

prokaryote cell numbers in all subsurface environments are of the order of 10^{30} (ref. 83). Although there are uncertainties in the estimates, and they are based on extrapolation from a small number of samples, deep subsurface microbial biomass may account for greater than 90% of global prokaryotic biomass, exceeding the values for all of the world's oceans and terrestrial environments by some margin.

Subsurface prokaryotic biomass may contain $3\text{--}5 \times 10^{17} \text{ g}$ of carbon, representing 60–100% of the carbon present in global plant biomass. Prokaryotic cells typically contain a higher proportion of nitrogen and phosphorus than plant biomass (about 10 times more), and thus the contribution of deep crustal microorganisms to global nitrogen and phosphorus pools is extremely significant⁸³. The deep biosphere is clearly of global significance simply in terms of key elemental budgets. However, the fact that many of the organisms present may be consuming and producing inorganic and organic compounds to generate energy and biomass extends their significance to the realms of biogeochemical processes. One such crucial process in the deep subsurface is the transformation of petroleum hydrocarbons to produce heavy oils by biodegradation.

of degradation is a major factor controlling the rate and extent of subsurface oil biodegradation, with diffusion of nutrients in the water leg probably being adequate. Biodegradation, like inorganic diagenesis in many petroleum reservoirs, may be isochemical⁷⁶, involving only mass transport of hydrocarbons, nutrients and oxidants largely within the reservoir²⁹. The system may operate even if closed at the reservoir scale, and water flow may not be needed, although undoubtedly it would help mineral dissolution and nutrient supply. Consequently, reservoir topology will be an important determinant of the rate and location of degradation. For example, in reservoirs filled by oil down to an under-seal (a shale for example) with limited water legs, the oil is frequently not extensively degraded, even when nearby oil columns with substantial water legs are heavily biodegraded. This may explain why degradation frequently stops before all reactive components are removed.

Hydrocarbon removal rates in biodegrading petroleum reservoirs

are of the order of $10^{-6} \text{ mmol oil l}^{-1} \text{ d}^{-1}$ in the degradation zone, which is comparable with many reported aquifer respiration rates^{29,77}. We think that these slow rates of degradation in nutrient-limited reservoirs are inadequate to support rapid resynthesis of the biochemicals necessary to support the activity of microorganisms at temperatures above 80 °C. This falls short of 113 °C, the current highest authenticated temperature at which life can be sustained⁷⁸, and is considerably lower than 121 °C, which has recently been reported as the new upper temperature for life⁷⁹. However, the majority of bacteria and archaea that are capable of growth at such high temperatures are isolated from environments rich in reduced electron donors, nutrients and electron acceptors, such as hydrothermal vent systems. Under such conditions, high metabolic rates required for the rapid regeneration of labile biomolecules can be sustained. This contrasts starkly with deep subsurface petroleum reservoirs, which are likely to be severely nutrient limited, and where low levels of metabolic activity appear to be the norm.

Box 3

Microorganisms in petroleum reservoirs

One of the first microbiological studies of a deep subsurface environment resulted in the isolation of sulphate-reducing bacteria from produced waters from an oil well³. Since that time a wide range of bacteria have been isolated from petroleum systems. These have exclusively come from samples of produced waters. Produced waters from reservoirs undergoing water injection to enhance oil recovery are prone to contamination from the injected waters, thus samples from non-water flooded reservoirs are more likely to yield organisms native to the petroleum reservoir, but even then non-indigenous organisms may be introduced during drilling and grow in pipework in the oil well. Consequently many aerobic heterotrophic bacteria, many of which are capable of hydrocarbon degradation only in the presence of oxygen, have been isolated from well head water samples. These organisms, however, are unlikely to be native populations from the petroleum reservoir. A wide range of anaerobic bacteria and archaea with physiological properties consistent with a deep subsurface existence have also been isolated (for an excellent review, see ref. 40). These include

fermentative thermophilic heterotrophic bacteria including *Thermotoga*, *Thermoanaerobacter* and *Fervidobacterium*. Hyperthermophilic archaea such as *Thermococcus* and *Archaeoglobus* have also been isolated from non-flooded reservoirs. In addition, sulphate-reducing bacteria, iron-reducing bacteria and methanogenic archaea have been identified in wellhead waters⁴². No bacteria capable of degrading hydrocarbons under *in situ* conditions have yet been isolated from petroleum reservoirs. Sediment samples have now been recovered from many deep subsurface environments. These have yielded isolates with characteristics that suggest that they are true representatives of the subsurface biosphere, and it has been possible to measure important biogeochemical processes in these sediments. Interestingly, there are no reports in the literature of successful isolation or characterization of the microbial communities in sediments recovered from a petroleum reservoir. The isolation and characterization of the deep slow biosphere in petroleum reservoirs thus remains a major challenge.

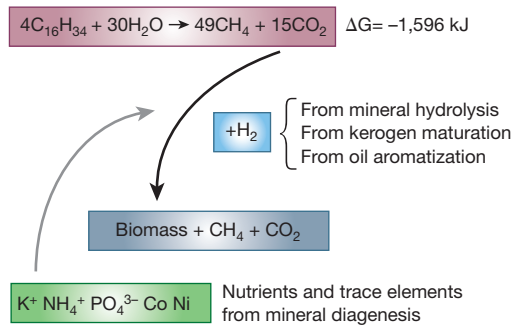


Figure 4 The putative chemistry of hydrocarbon degradation in most petroleum reservoirs with an absence of abundant sulphate. The overall conversion of hydrocarbons to biomass, methane and carbon dioxide may well involve water–hydrocarbon reactions^{47,49}, with the carbon dioxide produced being further reduced to methane using hydrogen produced either externally to or within the reservoir. Nutrient supply from mineral diagenesis may ultimately control the extent of hydrocarbon destruction and methane production^{69–72}. The detailed mechanisms of anaerobic hydrocarbon degradation are now being elucidated and many mechanisms involve addition of fumarate⁴⁸ or carboxylation of hydrocarbons to produce functionalized intermediates, which can either be metabolized to carbon dioxide or accumulate as dead-end metabolites. For example, a reductive pathway for the anaerobic degradation of naphthalene and 2-methylnaphthalene has been proposed by Annweiler and co-workers⁵⁰. Initial carboxylation of naphthalene is followed by degradation of 2-naphthoic acid and, then, via a series of hydrogenation steps, to 5, 6, 7, 8-tetrahydro-2-naphthoic acid. Further hydrogenation to octahydro-2-naphthoic acid may then be followed by further degradation steps to produce energy and carbon dioxide or by further hydrogenation to give decahydro-2-naphthoic acid (a possible dead-end metabolite). We have found these reduced naphthoic acid derivatives indicative of anaerobic hydrocarbon degradation in oils from all the biodegraded oil provinces that we have examined⁵¹.

Regeneration of heat-labile cell components might be an important mechanism permitting the growth of hyperthermophilic organisms at extremely high temperatures⁸⁰, and starvation survival of hyperthermophilic archaea is reduced at high temperatures⁸¹. Petroleum reservoir isolates are more resistant to starvation than similar isolates from hydrothermal vent systems. But when the geological timescales involved in the biodegradation of oil in high-temperature petroleum reservoirs are considered in conjunction with the extreme conditions, it may not be so surprising that bacteria and archaea in these environments are unable to survive at the upper limit for biological activity observed in other environments. We therefore suggest that low metabolic rate rather than biomacromolecule stability is the primary controller of the upper temperature limit of the petroleum-biodegrading deep subsurface biosphere. One can also speculate that this will be true of all life in oligotrophic deep crustal environments. The temperature base of the hydrocarbon-degrading biosphere may well be the base of life in Earth.

Perspective

Reservoir temperature and oil-charge histories are the primary controllers of the degree of oil biodegradation, with reservoir topology (relationships between oil and water volumes and interface areas) also exerting a control. Nutrient supply from temperature-dependent mineral diagenesis is a key factor, and methane is a major terminal product of oil biodegradation produced predominantly by carbon dioxide reduction with hydrogen. This has major implications both for conventional petroleum exploration and also for future strategies for recovery of heavy oil as methane through microbially engineered interventions.

The deep biosphere's origin and survival is intimately linked to the large scale tectonic processes that form, deform and destroy sedimentary basins. Observations of petroleum systems shed light on several key processes in deep biosphere evolution. Plate tectonics form and destroy sedimentary basins, accumulating organic carbon and converting it into mobile petroleum, which may migrate hundreds of kilometres before being consumed by microorganisms in transit or following accumulation in reservoirs. Mineral diagenesis and water flow in deep sediments control key nutrient supply to the petroleum biosphere, with low nutrient levels in petroleum reservoirs resulting in low metabolic rates and lower maximum temperatures for the survival of life. It appears from the distribution of biodegraded oils that temperatures near 80 °C represent the base of heterotrophic life in the crust, and this temperature may well be the base, in oligotrophic sediments, of all life in Earth. Uplifted reservoirs heated above 80–90 °C generally do not have biodegraded oils, indicating that they are effectively isolated from the surface over geological timescales, and thus the deep biosphere must be populated through burial processes, its organisms being inherited from the surface and sustained over geological timescales³⁵. □

doi:10.1038/nature02134

1. Roadifer, R. E. in *Exploration for Heavy Crude Oil and Natural Bitumen* (ed. Meyer, R.F.) 3–23 (Am. Assoc. Petrol. Geol., Tulsa, 1987).
2. Connan, J. in *Advances in Petroleum Geochemistry* Vol. 1 (eds Brooks, J. & Welte, D. H.) 299–335 (Academic, London, 1984).
3. Bastin, E. Microorganisms in oilfields. *Science* **63**, 21–24 (1926).
4. Krejci-Graf, K. Rule of density of oils. *Bull. Am. Assoc. Petrol. Geol.* **16**, 1038 (1932).
5. Bailey, N. J. L., Jobson, A. M. & Rogers, M. A. Bacterial degradation of crude oil: comparison of field and experimental data. *Chem. Geol.* **11**, 203–221 (1973).
6. Rubinstein, I., Strausz, O. P., Spyckerelle, C., Crawford, R. J. & Westlake, D. W. S. The origin of the oil sand bitumens of Alberta: a chemical and microbiological simulation study. *Geochim. Cosmochim. Acta* **41**, 1341–1353 (1977).
7. Alexander, R., Kagi, R. L., Woodhouse, G. W. & Volkman, J. K. The geochemistry of some biodegraded Australian oils. *Aus. Petrol. Explor. Assoc. J.* **23**, 53–63 (1983).
8. Volkman, J. K., Alexander, R., Kagi, R. L., Rowland, S. J. & Sheppard, P. N. Biodegradation of aromatic hydrocarbons in crude oils from the Barrow Sub-basin of Western Australia. *Org. Geochem.* **6**, 619–632 (1984).
9. Rowland, S. J., Alexander, R., Kagi, R. L., Jones, D. M. & Douglas, A. G. Microbial degradation of aromatic components of crude oils: A comparison of laboratory and field observations. *Org. Geochem.* **9**, 153–161 (1986).
10. Hunkeler, D., Jorger, D., Haberli, K., Hohener, P. & Zeyer, P. Petroleum hydrocarbon mineralisation in anaerobic laboratory aquifer columns. *J. Contaminant Hydrol.* **32**, 41–61 (1998).
11. Meredith, W., Kelland, S.-J. & Jones, D. M. Influence of biodegradation on crude oil acidity and carboxylic acid composition. *Org. Geochem.* **31**, 1059–1073 (2000).
12. Taylor, P. N., Bennett, B., Jones, D. M. & Larter, S. The effect of biodegradation and water washing on the occurrence of alkylphenols in crude oils. *Org. Geochem.* **32**, 341–358 (2001).
13. Wenger, L. M., Davis, C. L. & Isaksen, G. H. *Multiple Controls on Petroleum Biodegradation and Impact on Oil Quality*. (SPE 71450, Society of Petroleum Engineers, 2001).
14. Peters, K. E. & Fowler, M. G. Applications of petroleum geochemistry to exploration and reservoir management. *Org. Geochem.* **33**, 5–36 (2002).
15. Evans, C. R., Rogers, M. A. & Bailey, N. J. L. Evolution and alteration of petroleum in Western Canada. *Chem. Geol.* **8**, 147–170 (1971).
16. Hunt, J. M. *Petroleum Geochemistry and Geology* (Freeman, San Francisco, 1979).
17. Fedorak, P. M. & Westlake, D. W. S. Microbial degradation of alkyl carbazoles in Norman Wells crude oils. *Appl. Environ. Microbiol.* **47**, 858–862 (1984).
18. Zhang, C. et al. Effect of biodegradation on carbazole compounds in crude oils [in Chinese with English abstract]. *Shiyou yu tianranqi dizhi (Oil and Gas Geology)* **20**, 341–343 (1999).
19. Huang, H. P., Bowler, B. F. J., Zhang, Z. W., Oldenburg, T. B. P. & Larter, S. R. Influence of biodegradation on carbazole and benzocarbazole distributions in oil columns from the Liaohe basin, NE China. *Org. Geochem.* **34**, 951–969 (2003).
20. Mackenzie, A. S., Wolff, G. A. & Maxwell, J. R. in *Advances in Organic Geochemistry 1981* (eds Bjoroy, M. et al.) 637–649 (Wiley Heyden, London, 1983).
21. Thorn, K. A. & Aiken, G. R. Biodegradation of crude oil into non-volatile acids in a contaminated aquifer near Bemidji, Minnesota. *Org. Geochem.* **29**, 909–931 (1998).
22. Tomczyk, N. A., Winans, R. E., Shinn, J. H. & Robinson, R. C. On the nature and origin of acidic species in petroleum. I. Detailed acid type distribution in a California Crude Oil. *Energy Fuels* **15**, 1498–1504 (2001).
23. James, A. T. & Burns, B. J. Microbial alteration of subsurface natural gas accumulations. *Bull. Am. Assoc. Petrol. Geol.* **68**, 957–960 (1984).
24. Horstad, I. & Larter, S. R. Petroleum migration, alteration, and remigration within Troll Field, Norwegian North Sea. *Bull. Am. Assoc. Petrol. Geol.* **81**, 222–248 (1997).
25. Boreham, C. J., Hope, J. M. & Hartung-Kagi, B. Understanding source, distribution and preservation of Australian natural gas: A geochemical perspective. *Aus. Petrol. Explor. Assoc. J.* **41**, 523–547 (2001).
26. Larter, S. R. et al. When biodegradation preserves petroleum! Petroleum geochemistry of N. Sea Oil Rimmed Gas Accumulations (ORGAs). *Proc. AAPG Hedberg Research Conference on Natural Gas Formation and Occurrence, Durango, Colorado* (1999).
27. England, W. A., Mackenzie, A. S., Mann, D. M. & Quigley, T. M. The movement and entrapment of

- petroleum fluids in the subsurface. *J. Geol. Soc. Lond.* **144**, 327–347 (1987).
28. Barnard, P. C. & Bastow, M. A. in *Petroleum Migration* (eds England, W. A. & Fleet, A. J.) 167–190 (Spec. Publ. No. 59, Geol. Soc., London, 1991).
29. Larter, S. *et al.* The controls on the composition of biodegraded oils. (Part 1) Biodegradation rates in petroleum reservoirs in the deep subsurface. *Org. Geochem.* **34**, 601–613 (2003).
30. Masterson, W. D., Dzou, L. I. P., Holba, A. G., Fincannon, A. L. & Ellis, L. Evidence for biodegradation and evaporative fractionation in West Sak, Kuparuk and Prudhoe Bay field areas, North Slope, Alaska. *Org. Geochem.* **32**, 411–441 (2001).
31. Volkman, J. K., Alexander, R., Kagi, R. I. & Woodhouse, G. W. Demethylated hopanes in crude oils and their applications in petroleum geochemistry. *Geochim. Cosmochim. Acta* **47**, 785–794 (1983).
32. Peters, K. E. & Moldowan, J. M. *The Biomarker Guide* (Prentice Hall, New York, 1993).
33. Moldowan, J. M. & McCaffrey, M. A. A novel microbial hydrocarbon degradation pathway revealed by hopane demethylation in a petroleum reservoir. *Geochim. Cosmochim. Acta* **59**, 1891–1894 (1995).
34. Pepper, A. & Santiago, C. Impact of biodegradation on petroleum exploration and production: Observations and outstanding problems. *Abstracts of Earth Systems Processes*. (Geol. Soc. London, Geol. Soc. of America., Edinburgh, 24–28 June 2001).
35. Wilhelms, A., Larter, S. R., Head, I., Farrimond, P., di-Primio, R. & Zwach, C. Biodegradation of oil in uplifted basins prevented by deep-burial sterilisation. *Nature* **411**, 1034–1037 (2000).
36. Wellsbury, P., Goodman, K., Barth, T., Cragg, B. A., Barnes, S. P. & Parkes, R. J. Deep marine biosphere fuelled by increasing organic matter availability during burial and heating. *Nature* **388**, 573–576 (1997).
37. Winters, J. C. & Williams, J. A. Microbial alteration of crude oil in the reservoir. *Am. Chem. Soc. Div. Petr. Chem. Preprints* **14**, E22–E31 (1969).
38. Horstad, I., Larter, S. R. & Mills N. A quantitative model of biological petroleum degradation within the Brent Group reservoir in the Gullfaks field, Norwegian North Sea. *Org. Geochem.* **19**, 107–117 (1992).
39. Warren, E. A. & Smalley, P. C. *North Sea Formation Waters Atlas* (Geol. Soc. Lond., 1994).
40. Magot, M., Ollivier, B. & Patel, B. K. C. Microbiology of petroleum reservoirs. *Antonie Van Leeuwenhoek* **77**, 103–116 (2000).
41. Orphan, V. J., Taylor, L. T., Hafenbradl, D. & Delong, E. F. Culture-dependent and culture-independent characterization of microbial assemblages associated with high-temperature petroleum reservoirs. *Appl. Environ. Microbiol.* **66**, 700–711 (2001).
42. Röling, W. F. M., Head, I. M. & Larter, S. R. The microbiology of hydrocarbon degradation in subsurface petroleum reservoirs: perspectives and prospects. *Res. Microbiol.* **154**, 321–328 (2003).
43. Kartsev, A. A., Tabasarskii, Z. A., Subbota, M. I. & Mogilevski, G. A. in *Geochemical Methods of Prospecting and Exploration for Petroleum and Natural Gas* (eds Witherspoon, P. A., Romey, W. D.) 347 (Univ. California, California, 1959).
44. Connan, J., Lacrampe-Coulombe, G. & Magot, M. Origin of gases in reservoirs. *Proc. 1995 International Gas Research Conference, Government Institutes, Rockville, USA* 21–61 (1996).
45. Nilsen, R. K. & Torsvik, T. *Methanococcus thermolithotrophicus* isolated from North Sea oil field reservoir water. *Appl. Environ. Microbiol.* **62**, 1793–1798 (1996).
46. Nilsen, R. K., Beeder, J., Thorstenson, T. & Torsvik, T. Distribution of thermophilic sulfate reducers in North Sea oil field reservoir waters and oil reservoirs. *Appl. Environ. Microbiol.* **62**, 1793–1798 (1996).
47. Zengler, K., Richnow, H. H., Rossello-Mora, R., Michaelis, W. & Widdel, F. Methane formation from long-chain alkanes by anaerobic microorganisms. *Nature* **401**, 266–269 (1999).
48. Wilkes, H., Rabus, R., Fischer, T., Armstroff, A., Behrends, A. & Widdel, F. Anaerobic degradation of n-hexane in a denitrifying bacterium: Further degradation of the initial intermediate (1-methylpentyl) succinate via C-skeleton rearrangement. *Arch. Microbiol.* **177**, 235–243 (2002).
49. Widdel, F. & Rabus, R. Anaerobic biodegradation of saturated and aromatic hydrocarbons. *Curr. Opin. Biotechnol.* **12**, 259–276 (2001).
50. Annweiler, E., Michaelis, W. & Meckenstock, R. U. Identical ring cleavage products during anaerobic degradation of naphthalene, 2-methylnaphthalene, and tetralin indicate a new metabolic pathway. *Appl. Environ. Microbiol.* **68**, 852–858 (2002).
51. Aitken, C. M., Jones, D. M. & Larter, S. R. *Isolation and Identification of Biomarkers Indicative of Anaerobic Biodegradation in Petroleum Reservoirs*. (Geol. Soc. Lond., London, 2002).
52. Rueter, P. *et al.* Anaerobic oxidation of hydrocarbons in crude-oil by new types of sulfate-reducing bacteria. *Nature* **372**, 455–458 (1994).
53. Anderson, R. T. & Lovley, D. R. Biogeochemistry - hexadecane decay by methanogenesis. *Nature* **404**, 722–723 (2000).
54. Holba, A. G., Dzou, L. I. P., Hickey, J. J., Franks, S. G., May, S. J. & Lenney, T. Reservoir geochemistry of South Pass 61 field Gulf of Mexico: Compositional heterogeneities reflecting field filling history and biodegradation. *Org. Geochem.* **24**, 1179–1198 (1996).
55. Scott, A. R., Kaiser, W. R. & Ayers, W. B. J. Thermogenic and secondary biogenic gases, San-Juan Basin, Colorado and New Mexico - Implications for coalbed gas producibility. *Bull. Am. Assoc. Petrol. Geol.* **78**, 1186–1209 (1994).
56. Sweeney, R. E. & Taylor, P. Biogenic methane derived from biodegradation of petroleum under environmental conditions and in oil & gas reservoirs. *Proc. AAPG Hedberg Research Conference on Natural Gas Formation and Occurrence, Durango, Colorado* (1999).
57. Pallasser, R. J. Recognising biodegradation in gas/oil accumulations through the delta ¹³C composition of gas components. *Org. Geochem.* **31**, 1363–1373 (2000).
58. Dimitrakopoulos, R. & Meulenbachs, K. Biodegradation of petroleum as a source of ¹³C enriched carbon dioxide in the formation of carbonate cement. *Chem. Geol.* **65**, 283–289 (1987).
59. Barson, D., Bachu, S. & Esslinger, P. Flow systems in the Mannville Group in the east-Central Athabasca area and implications for steam-assisted gravity drainage (SAGD) operations for *in situ* bitumen production. *Bull. Can. Petrol. Geol.* **49**, 376–392 (2001).
60. Nazina, T. N., Ivanova, A. E., Borzenkov, I. A., Belyaev, S. S. & Ivanov, M. V. Occurrence and geochemical activity of microorganisms in high-temperature, water-flooded oil fields of Kazakhstan and Western Siberia. *Geomicrobiol. J.* **13**, 181–192 (1995).
61. Ng, T. K., Weimer, P. J. & Gawel, L. J. Possible nonanthropogenic origin of 2 methanogenic isolates from oil-producing wells in the San-Miguelito field, Ventura County, California. *Geomicrobiol. J.* **7**, 185–192 (1989).
62. Obratsova, A. Y., Shipin, O. V., Bezrukova, L. V. & Belyaev, S. S. Properties of the coccoid methylophilic methanogen, *methanococcoides-euhalobius* sp.-nov. *Microbiol.* **56**, 523–527 (1987).
63. Rozanova, E. P., Savvichev, A. S., Karavaiko, S. G. & Miller, Y. M. Microbial processes in the Savuiskoe oil-field in the Ob region. *Microbiology* **64**, 85–90 (1995).
64. Charlou, J. L., Donval, J. P., Fouquet, Y., Jean-Baptiste, P. & Holm, N. Geochemistry of high H₂ and CH₄ vent fluids issuing from ultramafic rocks at the Rainbow hydrothermal field (36 degrees 14' N, MAR). *Chem. Geol.* **191**, 345–359 (2002).
65. Chapelle, F. H. *et al.* A hydrogen-based subsurface microbial community dominated by methanogens. *Nature* **415**, 312–315 (2002).
66. Helgeson, H. C., Knox, A. M., Owens, D. H. & Shock, E. L. Petroleum, oil-field waters, and authigenic mineral assemblages – are they in metastable equilibrium in hydrocarbon reservoirs. *Geochim. Cosmochim. Acta* **57**, 3295–3339 (1993).
67. Manning, D. A. C. & Hutcheon, I. E. Distribution and mineralogical controls on ammonium in deep groundwaters. *Appl. Geochem.* (in the press).
68. Oldenburg, T., Huang, H., Donohoe, P., Willsch, H. & Larter, S. R. High molecular weight aromatic nitrogen and other novel hopanoid-related compounds in crude oils. *Org. Geochem.* (in the press).
69. Hiebert, F. K. & Bennett, P. C. Microbial control of silicate weathering in organic-rich ground-water. *Science* **258**, 278–281 (1992).
70. Rogers, J. R., Bennett, P. C. & Choi, W. J. Feldspars as a source of nutrients for microorganisms. *J. Am. Miner.* **83**, 1532–1540 (1998).
71. Bennett, P. C., Hiebert, F. K. & Rogers, J. R. Microbial control of mineral-groundwater equilibria: macroscale to microscale. *Hydrogeol. J.* **8**, 47–62 (2000).
72. Bennett, P. C., Rogers, J. R. & Choi, W. J. Silicates, silicate weathering, and microbial ecology. *Geomicrobiol. J.* **18**, 3–19 (2001).
73. Ehrenberg, S. N. & Jakobsen, K. G. Plagioclase dissolution related to biodegradation of oil in Brent Group sandstones (Middle Jurassic) of Gullfaks Field, Northern North Sea. *Sedimentol.* **48**, 703–721 (2001).
74. Whelan, J. K., Eglinton, L., Kennicutt, M. C. & Qian, Y. R. Short-time-scale (year) variations of petroleum fluids from the US Gulf Coast. *Geochim. Cosmochim. Acta* **65**, 3529–3555 (2001).
75. Yu, A. *et al.* How to predict biodegradation risk and reservoir fluid quality. *World Oil* **1–5** (April 2002).
76. Bjorlykke, K. in *Growth, Dissolution and Pattern Formation in Geosystems* (eds Jamtveit, B. & Meakin, P.) 381–404 (Kluwer, Netherlands, 1999).
77. Murphy, E. M. & Schramke, J. A. Estimation of microbial respiration rates in groundwater by geochemical modelling constrained with stable isotopes. *Geochim. Cosmochim. Acta* **62**, 3395–3406 (1998).
78. Blöchl, E., Rachel, R., Burggraf, S., Hafenbradl, D., Jannasch, H. W. & Stetter, K. O. *Pyrobolus fumarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. *Extremophiles* **1**, 14–21 (1997).
79. Kashefi, K. & Lovley, D. R. Extending the upper temperature limit for life. *Science* **301**, 934–934 (2003).
80. Stetter, K. O. Extremeophiles and their adaptation to hot environments. *FEBS Lett.* **452**, 22–25 (1999).
81. Takahata, Y., Hoaki, T. & Maruyama, T. Starvation survivability of *Thermococcus* strains isolated from Japanese oil reservoirs. *Arch. Microbiol.* **176**, 264–270 (2001).
82. Parkes, R. J. *et al.* Deep bacterial biosphere in Pacific Ocean sediments. *Nature* **371**, 410–413 (1994).
83. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
84. Demaison, G. J. Tar sands and supergiant oil fields. *Bull. Am. Assoc. Petrol. Geol.* **61**, 1950–1961 (1977).
85. Creaney, S. & Allan, J. in *Classic Petroleum Provinces* (ed. Brooks, J.) 189–202 (Geol. Soc., London, 1990).
86. Piggott, N. & Lines, M. D. in *Petroleum Migration* (eds England, W. A. and Fleet, A. J.) 207–226 (Geol. Soc., London, 1991).
87. James, K. H. in *Classic Petroleum Provinces* (ed. Brooks, J.) 9–36 (Geol. Soc., London, 1990).
88. Butler, R. M. Some recent developments in SAGD. *J. Can. Petrol. Technol.* **40**, 18–22 (2001).
89. Brooks, P. W., Fowler, M. G. & Macqueen, R. W. Organic geochemistry of Western Canada Basin tar sands and heavy oils. *Org. Geochem.* **12**, 519–538 (1988).
90. Riediger, C. L., Fowler, M. G., Snowdon, L. R., MacDonald, R. & Sherwin, M. D. Origin and alteration of Lower Cretaceous Mannville Group I oils from the Provost oil field, east Central Alberta, Canada. *Bull. Can. Petrol. Geol.* **47**, 43–62 (1999).
91. Rowland, S., Donkin, P., Smith, E. & Wraige, E. Aromatic hydrocarbon ‘humps’ in the marine environment: Unrecognized toxins? *Environ. Sci. Tech.* **35**, 2640–2644 (2001).

Acknowledgements Much of our work referenced in this insights article was funded and supported intellectually by two oil industry consortia, Phoenix (Norsk Hydro) and Bacchus (Shell, TexacoChevron, ExxonMobil, ConocoPhillips, Norsk Hydro, JNOC, Petrobras, Enterprise, TotalFinaElf, BP) and by NERC JREI awards. Specifically we acknowledge the contributions made to this article by our group at Newcastle and by colleagues elsewhere who have contributed comments and information. We acknowledge from Newcastle, B. Bennett, H. Huang, C. Aitken, A. Rowan, W. Röling, G. Love, P. Farrimond, A. Ross, B. Bowler, G. Rock, A. Aplin, K. Noke and from collaborating organizations, A. Wilhelms, M. Erdmann, H. Penteado, L. Trindade, L. Arauco, A. Murray, C. Riediger, S. Creaney, T. Dunn, L. Wenger, M. Li, M. Koopmans, M. Bowen, C. Zhang, J. Cody and M. Fowler for past contributions.

Fundamental principles and applications of natural gas hydrates

E. Dendy Sloan Jr

Center for Hydrate Research, Colorado School of Mines, Golden, Colorado 80401, USA (e-mail: esloan@mines.edu)

Natural gas hydrates are solid, non-stoichiometric compounds of small gas molecules and water. They form when the constituents come into contact at low temperature and high pressure. The physical properties of these compounds, most notably that they are non-flowing crystalline solids that are denser than typical hydrocarbons and that the gas molecules they contain are effectively compressed, give rise to numerous applications in the broad areas of energy and climate effects. In particular, they have an important bearing on flow assurance and safety issues in oil and gas pipelines, they offer a largely unexploited means of energy recovery and transportation, and they could play a significant role in past and future climate change.

Until mankind learns how to economically generate hydrogen for fuel cells, natural gas will be the premium fuel for this century for two reasons. First, gas burns cleanly, causes few pollution problems and, relative to oil or coal, produces less carbon dioxide. And second, liquid fuels are better used as feedstocks (raw material) for generation of petrochemicals. Two examples herald this coming change: many power plants are being converted from coal to natural gas, and fleets of cars have been converted from petrol to natural gas fuel.

As we deplete the readily accessible reserves, we will need to obtain natural gas from conditions that are both more severe and more remote. We will need to explore deep ocean environments with higher pressures, and permafrost environments with lower temperatures than we presently do. And gases that were previously thought to be uneconomical, such as those containing non-combustible components of nitrogen, carbon dioxide and hydrogen sulphide, will also be explored. Such unusual conditions also promote the formation of a solid compound of gas and water — namely clathrates of natural gas — commonly called gas hydrates.

Here, I indicate the motivation for hydrate science and engineering; that is, the applications where technical workers find a use for the physics, chemistry and biology that are associated with science. This is not to indicate that hydrate engineering is simply an applied science. As indicated recently in a defining book¹ on differences in technical philosophy, engineering frequently cannot afford the luxury of a thorough scientific foundation, and must proceed at risk. As only one example, the past decade's development of the new 'low-dosage' pipeline hydrate inhibitors proceeded in an Edisonian research mode (a process of intelligent guesswork, intuitive reasoning and testing), and scientific progress is currently being made to refine trial and error research gaps in inhibitor developments.

Below, I describe five major applications of hydrate research: flow assurance, safety, energy recovery, gas storage/transportation and climate change. Before the applications are addressed, an introduction to hydrate structures and their overall properties is presented. I conclude this review with an outline of future challenges. For readers who want a more detailed understanding, several hydrate reviews^{2–8} are available.

Hydrate structures

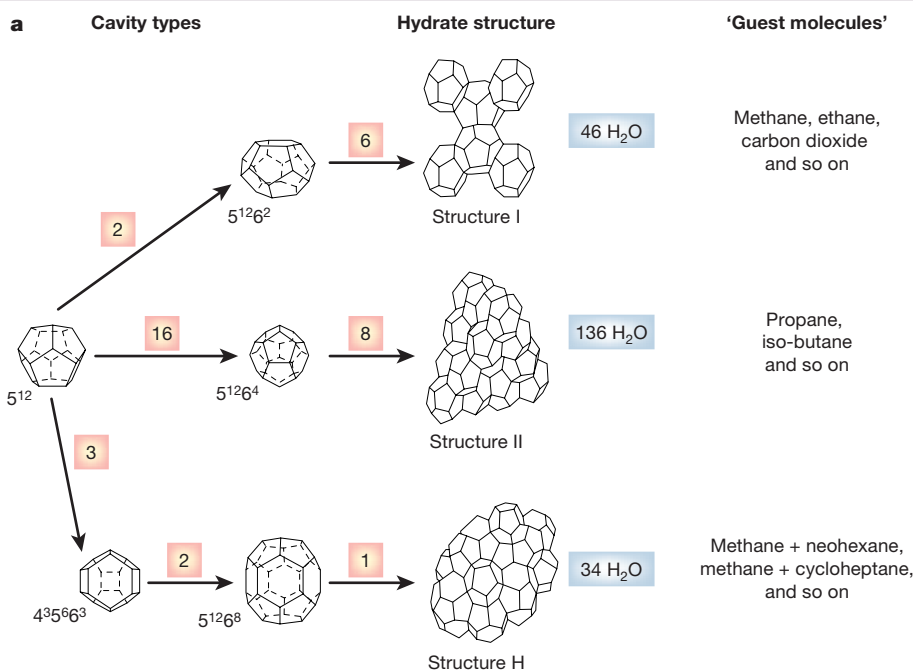
Clathrate hydrates typically form when small (< 0.9 nm) 'guest' molecules such as methane or carbon dioxide contact

water at ambient temperatures (typically less than 300 K) and moderate pressures (typically more than 0.6 MPa). On a molecular scale, single small guest molecules are encaged (enclathrated) by hydrogen-bonded water cavities in these non-stoichiometric hydrates. Guest-molecule repulsions prop open different sizes of water cages, which combine to form the three well-defined unit crystals shown in Fig. 1, and structural details and references have been provided in a recent book².

Cubic structure I predominates in the Earth's natural environments, and contains small (0.4–0.55 nm) guests; cubic structure II generally occurs with larger (0.6–0.7 nm) guests in mostly man-made environments; and hexagonal structure H may occur in either environment, but only with mixtures of both small and large (0.8–0.9 nm) molecules. The smallest hydrated molecules (Ar, Kr, O₂ and N₂), with diameters less than 0.4 nm, form structure II. Most hydrate science, and thus most applications, concentrates on structure I and structure II, with structure H in anecdotal occurrence. Although this review will emphasize structure I and structure II, many analogues occur for structure H hydrates.

Figure 1b lists the properties of the three common unit crystals. Water molecules form hydrogen bonds in a basic building block for both structure I and structure II, called the 5¹² (pentagonal dodecahedra) because there are 12 faces of pentagonally bonded water molecules in that cavity. Within the cavity, small guest molecules are enclathrated, with limited translation motion but substantially more rotation and vibration ability. The 5¹² building blocks are joined to other 5¹²s either through the vertices (structure I) or through the 5¹² faces (structure II).

Yet all structures need to fill space within their cavities to prevent hydrogen-bond strain and breakage. The 5¹² basic building blocks cannot fill space without bond breakage, and so interstices between the 5¹² cages are filled with other cavities that relieve the strain by incorporating hexagonal faces — two in the 5¹²6² cavity of structure I, and four hexagonal faces in the 5¹²6⁴ of structure II, in addition to the 12 pentagonal faces in each cavity. Thus, the cages can contain larger guest molecules. The cages form basic repeating unit crystals with ratios of 2•5¹² + 6•5¹²6² in structure I and 16•5¹² + 8•5¹²6⁴ in structure II, indicating sixteen 5¹²s and eight 5¹²6⁴s in the structure II unit crystal. Although both large and small cages are present in each crystal structure, sometimes single guests are too large for the smaller cage, which must go empty so that only the larger cage is filled. However, smaller molecules can fill both cages.



b

Hydrate crystal structure	I		II		H		
Cavity	Small	Large	Small	Large	Small	Medium	Large
Description	5 ¹²	5 ¹² 6 ²	5 ¹²	5 ¹² 6 ⁴	5 ¹²	4 ³ 5 ⁶ 6 ³	5 ¹² 6 ⁸
Number of cavities per unit cell	2	6	16	8	3	2	1
Average cavity radius (Å)	3.95	4.33	3.91	4.73	3.91 [†]	4.06 [†]	5.71 [†]
Coordination number*	20	24	20	28	20	20	36
Number of waters per unit cell	46		136		34		

*Number of oxygens at the periphery of each cavity.

[†]Estimates of structure H cavities from geometric models.

Figure 1 The three common hydrate unit crystal structures. Nomenclature: 5¹²6⁴ indicates a water cage composed of 12 pentagonal and four hexagonal faces. The numbers in squares indicate the number of cage types. For example, the structure I unit crystal is composed of two 5¹² cages, six 5¹²6² cages and 46 water molecules.

In all three structures, typically there is only one guest molecule within each cage. However, at unusual conditions such as very high pressure, it is possible to have multiple-cage occupancy with unusually small guests, such as hydrogen or noble gases. For example, it was recently shown⁹ that hydrogen can form hydrates at very high pressure with as many as two occupants in the small cage and four in the large cage of hydrate structure II. However, very small guests and multiple occupancies are considered an aberration.

Remarkably, when all hydrate cavities are filled, the three crystal types have similar concentrations of components: 85 mol% water and 15 mol% guest(s). Hydrate formation is most likely to take place at the interface between the bulk guest and aqueous phases, because the hydrate component concentrations greatly exceed the mutual fluid solubilities. The solid hydrate film at the interface acts as a barrier to prevent further contact of the bulk-fluid phases, and fluid surface renewal is required for continued clathrate formation.

Guest to cage ratio

Examination of the size ratios of the guest molecules to the cages they occupy can aid the understanding of hydrates. This heuristic controls not only the occupancy but also the thermodynamic properties of these structures, to a first approximation.

Simple hydrate guests

A number of researchers have commented on the fit of the guest molecule within the hydrate cage, beginning with von Stackelberg¹⁰, whose modified original diagram is shown in Fig. 2. Table 1 lists the

size ratios of guests of natural gas in the four common cavities of structure I and structure II. The discriminating size ratio is not absolute for each cage, instead it occurs over a molecular size range, which has a number of important implications. The implication is that clathrate hydrates have no fixed size ratio of guest to host, as shown by the ranges in Fig. 2. It is particularly interesting to note the resulting behaviour of molecules at cage size borders.

Take, for example, the most common natural gas compound, methane, in the phase diagram for methane and water in Fig. 3 (ref. 11). The non-stoichiometry of the hydrate causes the hydrate composition area (vertical axis parabola) shown in green in Fig. 3 rather than the vertical stoichiometric line originally proposed¹² for methane hydrate. Assuming equilibrium, the implication of this hydrate phase area is that laboratory-made hydrates (from methane-rich systems) may differ slightly in composition from *in situ* hydrates, which can form in methane-lean systems. This methane hydrate composition difference, although small (4%), when multiplied by the entire hydrate reserve is sufficient to supply the USA for 600 years at the current energy usage.

Binary hydrate guests

For binary systems, Holder and Manganiello¹³ indicated that an optimized fit of guests in the cages was sufficient for hydrate azeotropes, for which the vapour composition is equal to that of the hydrate. This is particularly unusual because azeotropy in vapour-liquid equilibria is only possible for a non-ideal solution (with an activity coefficient), whereas hydrates typically form ideal solid solutions.

The earlier measurements of von Stackelberg¹⁰ showed that some single guest molecules that are structure I formers will form structure II in binary guest mixtures — a somewhat counter-intuitive notion. This idea was later predicted¹⁴ and measured¹⁵ over a wide composition range for methane and ethane mixtures. However, other structure I simple guest combinations such as methane and carbon dioxide will not form stable structure II as binary mixtures¹⁶. The question arises as to why this occurs for methane and ethane mixtures but not for methane and carbon dioxide mixtures.

The reason for such structural transitions was addressed initially

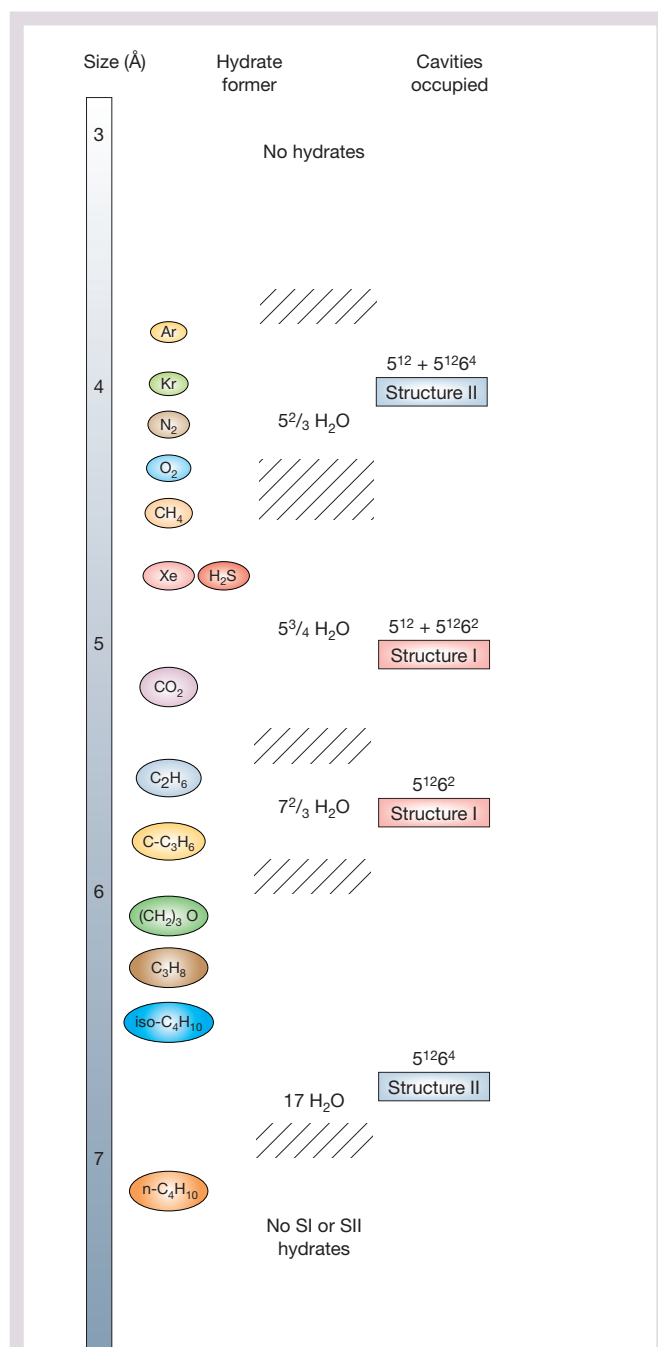


Figure 2 Hydrate guests versus hydrate cavity size ranges. Along the line are the size of the guest molecules in hydrates. The broad shaded areas and the numbers on the right discriminate the number of water molecules in hydrates occupied by single guest occupants shown on the left. For example, methane has a typical hydration number of 5 $\frac{1}{4}$ and occupies both the 5¹² and 5¹²6² cavities of structure I. However, propane is so large, it can only fit into the largest structure II hydrate cavity (5¹²6⁴). Adapted from ref. 43. Copyright Geological Society, London.

by Ripmeester³ and subsequently by Hester and Sloan¹⁷. The structure II transition occurs when two molecules are at each extreme of the structure I molecular sizes shown in Fig. 2. That is, small structure I formers in the 5¹² cage (which are almost small enough to form structure II) and large structure I formers in the 5¹²6² cage (which are almost large enough to form structure II) will, when mixed, cause structure II to form from two structure I formers.

Two other examples serve to indicate that hydrate structure research has been a minefield for the complacent who think that everything structural is known. First, for almost 40 years, until the prediction of structure II formation¹³ by the smallest guests was validated by Davidson *et al.*¹⁸, it was erroneously thought that they formed structure I as simple hydrates. Second, in 1987, structure H was discovered¹⁹, but on close inspection structure H crystals had been measured much earlier in the diffraction data of von Stackelberg¹⁰. Again for four decades, the hydrate community paradigm forced the data into structure I and structure II categories, counting structure H as an aberration.

Physical properties and implications

Several key physical properties of hydrates determine the roles that they play (or might play in the future) in both industry and the environment. They are solids with densities greater than those of typical fluid hydrocarbons, and this has practical implications for flow assurance in pipelines and the safety thereof. Furthermore, the fact that, in effect, hydrates concentrate their guest molecules results in three potential applications: that energy can be recovered from *in situ* hydrates; that hydrates can be used to transport stranded gas; and that hydrates may be a factor in climate change. Each of these implications and applications is briefly discussed.

Flow assurance

First, and perhaps most importantly, when hydrates form, they are solid, non-flowing crystalline structures. This leads to the most urgent consideration — that of flow assurance in oil and gas pipelines. Oil and gas wells always produce undesired water along with hydrocarbons that are in the hydrate guest size range. As the flowing mixed phases cool, hydrates form and plug transmission lines, causing costly production stoppages, sometimes for as long as months, in large pipelines, while the hydrates are dissociated.

The petroleum industry would like to maintain their processes outside the hydrate stability range. Fortunately, the hydrate stability temperature and pressure range is predictable to within experimental accuracy using modern thermodynamic programs usually based upon a Gibbs energy extension²⁰ of the van der Waals and Platteeuw²¹ method. Unfortunately, however, low temperatures (such as the deep-sea floor temperature of 277 K) and the mandates of high pressure for economic energy densities place many pipelines well within the hydrate-formation region. High pressures and low temperature require hydrate-inhibition methods. During the past decade this phenomenon has initiated a new type of engineer — the flow-assurance engineer — whose major objective is to prevent pipeline blockages,

Table 1 Ratios of molecular diameters* to cavity diameters† for some molecules including natural gas-hydrate formers

Molecule	Guest diameter (Å)	Structure I		Structure II	
		5 ¹²	5 ¹² 6 ²	5 ¹²	5 ¹² 6 ⁴
N ₂	4.1	0.804	0.700	0.817 ^F	0.616 ^F
CH ₄	4.36	0.855 ^F	0.744 ^F	0.868	0.652
H ₂ S	4.58	0.898 ^F	0.782 ^F	0.912	0.687
CO ₂	5.12	1.00	0.834	1.02	0.769
C ₂ H ₆	5.5	1.08	0.939 ^F	1.10	0.826
C ₃ H ₈	6.28	1.23	1.07	1.25	0.943 ^F
i-C ₄ H ₁₀	6.5	1.27	1.11	1.29	0.976 ^F
n-C ₄ H ₁₀	7.1	1.39	1.21	1.41	1.07

^F Indicates the cavity occupied by the single hydrate guest.

*Molecular diameters obtained from von Stakelberg¹².

†Cavity radii from Table 2-1 minus 1.4 Å water radii.

primarily of hydrates, but also of waxes and other solids.

To provide flow assurance, the energy industry injects hydrogen-bonding fluids (for example, alcohols or glycols) into the flowing stream at the wellhead to compete with solid hydrates for the available water. Worldwide methanol costs for hydrate inhibition are estimated at US\$220 million annually (P. K. Notz, personal communication). In addition, severe financial penalties are paid for large methanol storage capacity on offshore platforms and for greater than 50 p.p.m. methanol contaminations in refinery feedstocks. During the production process from hydrocarbon reservoirs, any amount of water, which generally increases over the life of the well, must be hydrate inhibited. Although the aqueous phase is the primary place where hydrate inhibition occurs, a small concentration (typically less than 0.1 mol%) of methanol partitions into the hydrocarbon vapour and liquid phases.

Yet the bulk of methanol is lost to the hydrocarbon phases because, even with low methanol concentrations, the hydrocarbon phase amount greatly exceeds that of the aqueous phase. Methanol is a difficult molecule to model partitioning into both hydrocarbon liquid and vapour phases, particularly with high accuracy, at low concentrations. If we apply Hildebrand's solubility maxim, 'like dissolves like', we conclude that the methyl group makes methanol soluble in hydrocarbons, whereas the hydroxyl group increases its solubility in water. Thus any equation-of-state that models methanol partitioning must address the challenge of modelling hydrocarbon and aqueous phases equally well.

In addition to hydrogen-bonding inhibitors, which require high concentrations, new low-dosage hydrate inhibitors — both kinetic inhibitors and anti-agglomerants — have been developed in the past decade to prevent hydrate crystal growth and agglomeration, respectively. These new chemicals are rapidly being adopted in the field²² and provide a fertile research area for molecular modelling.

Safety

Secondly, the hydrate solid specific gravity is typically 0.9 compared with typical fluid hydrocarbon specific gravities of 0.8 or less. This higher density leads to the problem of ensuring hydrate safety and preventing the annual loss of property and lives. When hydrate blockages dissociate in pipelines, they detach first at the pipe wall; therefore, any pressure gradient across the high-density hydrate plug will cause the hydrate to travel rapidly (measured at 300 km hr⁻¹) down the pipeline. This effect will compress the downstream gas, either causing pipeline blowouts or causing the plug to erupt through pipeline bends. Case studies of hydrate safety problems are given in a monograph on this subject⁴.

A second safety concern arises when hydrate plugs are locally heated (for example, using a blowtorch outside a pipeline) to dissociate them. Frequently, the evolving gas from the hydrate is contained by the ends of the plug until the pipeline bursts owing to the pressure being too high. This safety concern is a result of the next hydrate property — the ability of hydrates to concentrate high levels of gas in a hydrated form.

Energy recovery

When small (≤ 0.9 nm) hydrocarbon guest molecules are encaged in hydrates, with typically one molecule per cavity, the guests are separated approximately 0.5 nm by the water cages. This means that the energy density in hydrates is approximately the same as that of a compressed gas, that is, less than the energy density of liquefied natural gas (LNG). For example, if every hydrate cavity were filled with a guest molecule, one volume of hydrate would dissociate to 180 volumes (STP) of gas. The gas concentration in clathrates is comparable to that of a highly compressed gas (that is, methane gas at 273 K and 18 MPa).

A large fraction of the Earth's fossil fuels is stored in clathrate hydrates. Even the most conservative current estimates²³ suggest that the amount of energy in hydrates is equivalent to twice that of all

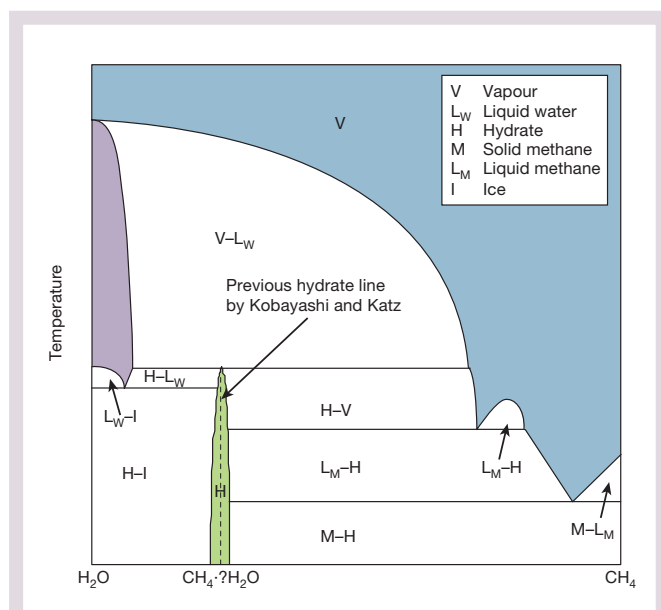


Figure 3 The isobaric methane and water phase diagram. Compare the vertical parabolic hydrate area (green) with the previous vertical stoichiometric hydrate line of Kobayashi and Katz¹². The parabolic region is a result of incomplete filling of the small cages (5¹²) in structure I. Variation in hydrate cage filling and the resulting hydrate parabola is a function of temperature, overall methane composition and pressure (not shown here).

other fossil fuels combined. Most of the natural-gas-containing hydrates are in the ocean bottom, and although production of gas from such deep-lying hydrates is now too expensive, it is likely that within the next two decades we will tap that fuel source to meet growing energy demands. Table 2 compares hydrated methane to that in conventional reserves for 11 arbitrary divisions of the world.

Most of the natural hydrates around the world are biogenic — the guest gas comes from bio-degraded plant and animal matter that have been buried in the sea floor at low temperature over long periods. Substantial but anecdotal evidence exists for thermogenic hydrates from deeper gas sources in places like the Gulf of Mexico²⁴ and the Caspian Sea²⁵. Most of the estimates of gas hydrates have come from indirect seismic evidence using a bottom simulating reflector (BSR), which indicates reflections from gas at the base of the hydrate (see Fig. 4).

BSR indications are not totally reliable, and other more accurate methods are needed. There are a significant number of cases in which hydrates occurred without bottom simulating reflections, or when the BSR did not indicate the presence of hydrates. Notwithstanding this problem, the resource numbers are so large that they warrant energy-recovery studies, even if they are in error by as much as two orders of magnitude.

Much of the available public funding of hydrate research has been channelled toward industrial field experiments, aimed at the production of energy. Although results from industrial experiments (for example, in Alaska and in Japan) may not be publicly available, results from two recent drilling expeditions are soon to be published. These detailed field experiments will probably serve as design bases for the foreseeable future, owing to their thorough documentation.

Pilot drilling, characterization and production testing of hydrates have begun in permafrost regions, which have higher concentrations of hydrates (for example, 30 vol.% in the 1998 Mallik 2L-38 well in Canada), to learn how to approach the more dispersed, but much greater, ocean resource in the future. The Mallik 5L international field experiment was concluded in March 2002 on Richards island in the MacKenzie delta of Canada at a cost of US\$17 million. This permafrost experiment provided the first direct evidence that hydrates

could be economically recovered at high concentrations. The information provided by this experiment, when it becomes available in a Geological Survey of Canada report early in 2004 (ref. 26), will be a landmark upon which the industry will base designs for energy recovery.

For ocean hydrates, the benchmark will be Leg 204 of the Ocean Drilling Program (ODP), completed in September 2002 on Hydrate ridge of the Cascadia Margin off Oregon. This is one of the first ocean drillings for hydrates to recover large, pervasive hydrates, other than anecdotal evidence. Much of the information from Leg 204 was presented in preliminary findings at the combined American Geological Union, European Geophysical Society, European Union of Geosciences meeting in Nice (April 7–10, 2003). This drilling, the result of which should appear in print in the final quarter of 2003, provides the first evidence that ocean hydrates may be present in sufficient concentration to be economically producible.

Until the publication of the benchmark results from the Mallik 5L and Leg 204 wells, the best literature for natural hydrates can be found in summaries and volumes about the Mallik 2L well²⁷, the Blake Ridge ODP²⁸ Leg 164 and the Gulf of Mexico²⁴.

It should be noted that the amount of energy in ocean hydrates is several orders of magnitude greater than that in permafrost hydrates. Put another way, the error in the ocean hydrated energy estimate is greater than the entire permafrost hydrated energy estimate. Until ODP Leg 204, however, it was thought that the most concentrated hydrates were in the permafrost, which provided more accessible recovery.

Methods for the economic recovery of methane from natural hydrates are uncertain, and substantial creativity has gone into devising new recovery methods, as well as into applying existing oil and gas technology to hydrate recovery. All recovery methods apply one or more of the following three principles: (1), reduction of the pressure below that of hydrate stability; (2), addition of enough energy to disrupt the water hydrogen bonds; and (3), addition of strong hydrogen-bonding chemicals (such as alcohol or glycol) to disrupt the hydrate structure at reservoir conditions.

Finite difference reservoir recovery models²⁹ indicate that gas production is only economical at rates larger than 500,000 standard cubic metres (0.1 MPa and 289 K) per day. This will require both depressurization and thermal/inhibitor stimulation. The most producible of the permafrost hydrate deposits are those lying adjacent to

a gas reservoir, because free gas production will dissociate hydrates by decreasing reservoir pressures below hydrate stability. Heat from the Earth allows hydrate decomposition to slowly replenish the gas reservoir. Makogon⁵ indicated that a Siberian permafrost reservoir was produced in this manner during the 1970s.

Gas production from hydrates close to conventional permafrost reservoirs will begin in the West during the next decade at incremental costs over normal gas production. Production from stand-alone hydrates in the permafrost or in the ocean will be much more costly, but is technically feasible. Both Japanese and American programmes forecast that stand-alone hydrated energy recovery will begin by 2015.

Storage and transportation

It is estimated that about 70% of the total gas reserve is either too far from an existing pipeline or too small to justify a liquefaction facility. Gudmundsson and Borrehaug³⁰ suggested that it is economically feasible to transport stranded gas in hydrated form. In the fourth international hydrate conference, workers from Mitsui Shipbuilding³¹ showed that work in conjunction with the Japan Maritime Research Institute³² provides a basis for extending the basic concept of Gudmundsson *et al.*³³ to transport stranded gas.

Climate change

A recent publication³⁴ thoroughly documents evidence for Late Quaternary climate change caused by hydrates, commonly called 'the hydrate gun hypothesis'. The concept is that, as little as 15,000 yr ago, methane from hydrates caused significant global warming.

The hydrate gun hypothesis seems analogous to another, somewhat less controversial, hypothesis, proposed by Dickens *et al.*^{35–37}. They suggested that an ancient (55.5 Myr ago), massive ocean methane hydrate dissociation might explain a 4–8 °C temperature rise over a brief geological time interval (10³ years) called the Late Palaeocene Thermal Maximum (LPTM). This is documented in deep ocean drilling samples as a prominent negative carbon isotope ($\delta^{13}\text{C} = -2.5\text{‰}$) in ocean sediments, in fossil tooth enamel, and in carbonates and organic sediments in terrestrial sequences. This $\delta^{13}\text{C}$ reduction in the ocean and the recovery over the ensuing 0.2 million years (see Fig. 5a) is consistent with pronounced dissolution of calcium carbonate in the deep sea sediment deposited during the LPTM, as shown in Fig. 5b.

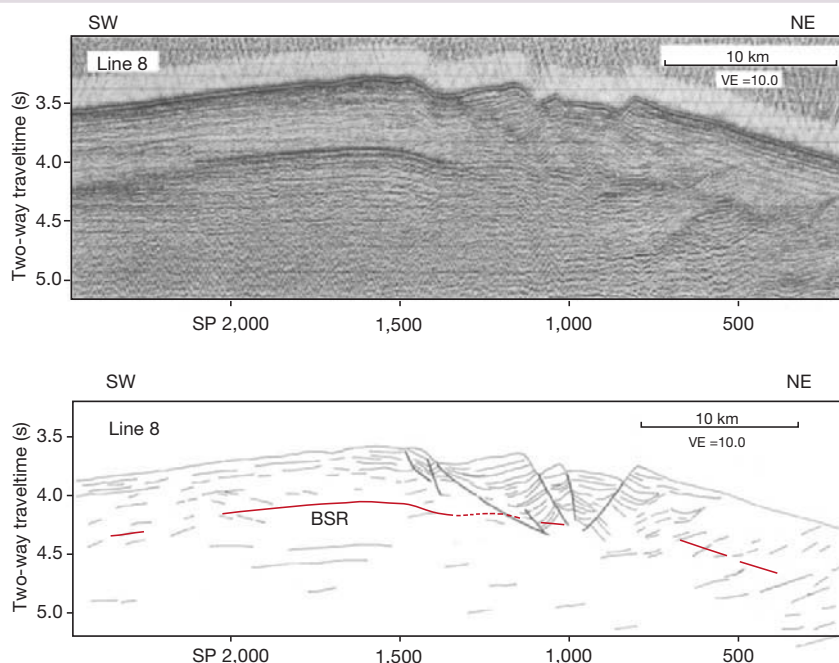


Figure 4 A seafloor slump in the Blake-Bahama ridge shown in both seismic (top) and cartoon (bottom) relief⁴⁰. Note the bottom simulating reflector (BSR) parallel to the ocean bottom, except in the middle section (dotted line) where it appears that a seafloor eruption has occurred. Reproduced with permission from ref. 40. Copyright Geological Society, London.

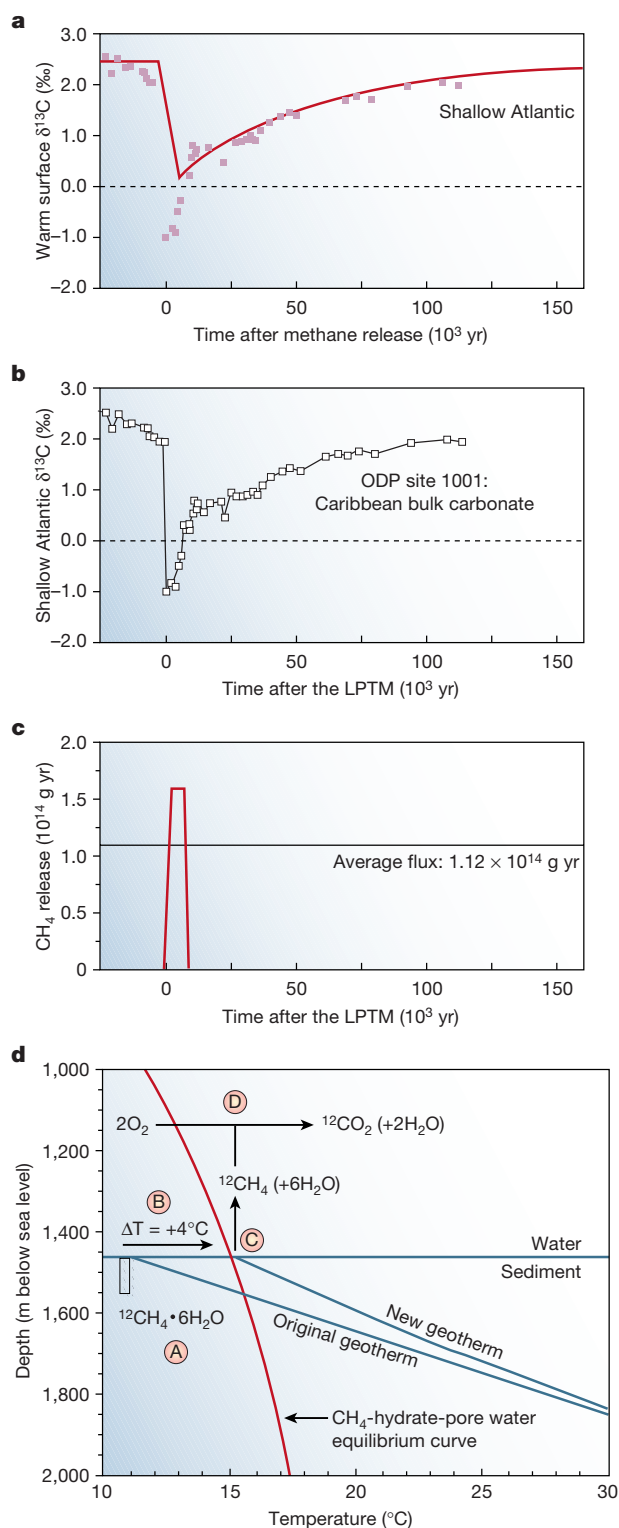


Figure 5 Hypothesized causes of the Late Paleocene Thermal Maximum (LPTM). **a**, Carbon isotope reduction and recovery during LPTM. **b**, Dissolution of calcium carbonate during LPTM. **c**, Evolution of methane from hydrates. **d**, Initial shift in ocean hydrate equilibrium curve to cause the methane release. In sequence: in **d**, the geotherm shifted by 4°C , causing release of a large quantity of methane from hydrates, shown in **c**. The result was a $\delta^{13}\text{C}$ isotope reduction and recovery, as shown in **a** and **b** through conversion of methane, first to carbon dioxide and then to calcium carbonate. Adapted from ref. 44. Copyright Société géologique de France.

Table 2 Conventional and hydrated gas resources in trillion cubic metres

Region	Conventional gas (TCM)	Methanehydrate (TCM)
North America	32.82	6,853
Latin America and Caribbean	21.1	5,139
Western Europe	15.27	856
Central and Eastern Europe	2.05	0
Former Soviet Union	117	4,711
Middle East and North Africa	77.2	214
Sub-Saharan Africa	13.9	429
Central Asia and China	10.07	429
Pacific OECD	2.68	1,713
Other Pacific Asia	11.18	214
South Asia	4.72	429
Total	310.3	20,987

TCM, trillion cubic metres

In the LPTM hypothesis, the evolution of a large amount of methane from hydrates (1.12×10^{18} g of methane) is the only plausible explanation that has been offered to explain this environmental perturbation. The abnormal $\delta^{13}\text{C}$ isotope indicates that the source was external to the normal ocean-atmospheric-biomass carbon pool. Figure 5c shows a rapid evolution of methane from hydrates; methane is hypothesized to be oxidized to carbon dioxide that is greatly enriched in $\delta^{12}\text{C}$.

Figure 5d shows the hydrate equilibrium curve as a function of depth and temperature in the ocean. Hydrates are only stable between the equilibrium line and the original geotherm to the left of the curved line, at depths below the sediment surface, shown by the small vertical rectangle at A. In the LPTM, if the ocean was warmed by 4°C , the hydrates between the original geotherm and the equilibrium curve would melt, as the new geotherm was established. The warming from the original to the new geotherm would result in methane expulsion to the environment, where it would be oxidized to carbon dioxide, resulting in significant further warming. It was hypothesized that the resulting carbon dioxide was re-absorbed by the ocean over the ensuing 0.2 Myr.

The importance of the LPTM perturbation is that it is the first well documented instance of an explanation for how the global carbon cycle and other systems relate to a rapid, massive input of fossil fuel, such as may occur in modern industrial times. The data and summary in the publication by Kennett³⁴ are the most thorough source of information in support of extending the theory to more modern times (the Late Quaternary), about 15,000 yr ago. However, there is a considerable controversy concerning the validity of the hypothesis, as suggested below.

In the recent meeting, it was suggested³⁸ that 'the hydrate gun is firing blanks', and that the atmospheric methane spike was due to emissions from wetlands and peat bogs. This new theory requires a glacial-interglacial vegetation time shift of 1,000 Gt C, which the proposers of the theory, Maslin and Thomas, admit is difficult. However, even this counter hypothesis requires some hydrate-derived methane for a mass balance, along with a shift in time for the wetlands.

In a review of the hydrate gun monograph³⁴, Dickens³⁹ generally concurs with the theory, but criticizes it on the grounds that it "perpetuates the common misconception that present-day methane hydrates are stable. These systems may be in a steady state, but they must be viewed as dynamic, with large carbon fluxes to and from the ocean, even at [the] present day".

In closing the discussion on hydrate-related climate change, it should be noted that seafloor hydrate dissociation is also directly related to slumping of sediments on the sea floor. Significant hydrated sediment slumps in the ocean can jeopardize the foundation of sub-sea structures, such as platforms, manifolds and pipelines. The single incident off the Carolina coast shown in Fig. 4 took place about 15,000 years ago⁴⁰ and increased the extant Earth's atmospheric methane by as much as 4%. The effect of subsidence on sub-sea structures and foundations represents the initial meeting point for the two

energy communities — the first is concerned with hydrates in the Earth, and the second with concerns for hydrates in man-made production systems. The interested reader is referred to the recent monograph⁴¹ on this topic.

Future challenges

Currently, it appears that hydrate research has acceptably addressed the thermodynamic challenge for most conditions. Time-independent hydrate quantification, courtesy of extensions²⁰ of the van der Waals and Platteeuw²¹ model, is at the bounds of experimental accuracy, and is the most common industrial exemplar of statistical thermodynamics use.

The most accurate thermodynamics have been provided through modern spectroscopy for hydrate phase measurements, by Raman, NMR and diffraction spectroscopy, through the bridge of statistical thermodynamics to the macroscopic domain. The incorporation of these three spectroscopic methods have enabled more accurate descriptions of hydrate mixtures, so that mixture thermodynamic predictions no longer depend solely on single hydrate guest measurements. An overview of such hydrate spectroscopic methods and results is provided in a recent review⁶.

Although the central concerns of hydrate thermodynamics have been addressed, challenges remain at the periphery — for example, at very high pressures (> 1,000 bar), in unusual fluids such as black oils, hydrate–sediment mixtures, and the methanol-partitioning challenge indicated earlier. As an example of one such challenge, hydrate–sediment mixtures have an unexplained thermal diffusivity maximum when plotted against sediment concentration. As we begin to examine hydrates in nature, such challenges for time-independent properties will require decades to resolve.

However, the largest challenge is to describe the kinetics of hydrates⁴². The fact that hydrates are solid compounds makes their slow, solid-phase kinetics particularly challenging to researchers. An additional challenge arises from the fact that hydrate solids form interfacial barriers between the liquid and vapour phases that typically compose them. Hydrate research is most accurate when studying a time-independent target. Typically, time-dependent (kinetic) research is much more difficult and at least an order of magnitude of accuracy is lost, relative to time-independent (thermodynamic) research.

The use of kinetic-model results to predict data from other laboratories is problematic. Molecular dynamic simulations of hydrate kinetics have been hindered by stochastic nucleation and the large number of molecules and time required for growth processes. More hydrate phase measurements are required to provide a needed breakthrough — a unified hydrate kinetics model.

Conclusions and outlook

Wherever small molecules contact water, the potential for a hydrate phase should be considered. The size ratio (guest to cavity) determines hydrate structural stability to a first-order approximation. Other simple hydrate properties such as solid behaviour, density and concentration of guest molecules affect the major applications of hydrate safety, flow assurance, energy production and storage and climate change.

During the next decade, gas production will begin from permafrost hydrates associated with conventional gas reservoirs. However, efficient production of ocean hydrates is problematic and requires an engineering breakthrough to be economically feasible. Yet, the potential to tap the Earth's largest hydrocarbon energy resource cannot be ignored.

Although hydrate thermodynamics are understood to an acceptable degree for most engineering applications, the kinetics arena will represent the largest challenge for advancing the information on hydrates. Although we know quite a lot about what hydrates are, the question of how hydrates form is still very much unanswered. Find-

ing the answers to such questions provides the intrinsic motivation for future research.

doi:10.1038/nature02135

- Koen, B. V. *Discussion of the Method: Conducting the Engineer's Approach to Problem Solving* (Oxford Univ. Press, Oxford, 2003).
- Sloan, E. D. in *Clathrate Hydrates of Natural Gases* (Marcel Dekker, New York, 1998).
- Ripmeester, J. A. Hydrate research — from correlations to a knowledge-based discipline: the importance of structure. *Ann. NY Acad. Sci.* **912**, 1–16 (2000).
- Sloan, E. D. in *Hydrate Engineering* (Soc. Petrol. Eng., Richardson, TX, 2000).
- Makogon, Y. F. *Hydrates of Hydrocarbons* (Pennwell Books, Tulsa, 1997).
- Sloan, E. D. Clathrate hydrate measurements: microscopic, mesoscopic, and macroscopic. *J. Chem. Therm.* **35**, 41–53 (2003).
- Davidson, D. W. in *Water: A Comprehensive Treatise* (ed. Franks, F.) (Plenum, New York, 1973).
- Englezos, P. Clathrate hydrates. *Ind. Eng. Chem. Res.* **32**, 1251–1274 (1993).
- Mao, W. L. *et al.* Hydrogen clusters in clathrate hydrate. *Science* **297**, 2247–2249 (2002).
- von Stackelberg, M. Solid gas hydrates. *Naturwissenschaften* **11–12**, 1–22 (1949).
- Huo, Z., Hester, K., Miller, K. T. & Sloan, E. D. Methane hydrate non-stoichiometry and phase diagram. *Am. Ind. Chem. Eng. J.* **49**, 1300–1306 (2003).
- Kobayashi, R. & Katz, D. L. Methane hydrate at high pressure. *J. Petrol. Technol.* **2579**, 66–70 (1949).
- Holder, G. D. & Manganiello, D. J. Hydrate dissociation pressure minima in multicomponent systems. *Chem. Eng. Sci.* **37**, 9–16 (1982).
- Hendriks, E. M., Edmonds, B., Moorwood, R. A. & Szczepanski, R. Hydrate structure stability in simple and mixed hydrates. *Fluid Phase Equilibria* **117**, 293–298 (1996).
- Subramanian, S., Kini, R. A., Dec, S. F. & Sloan, E. D. Evidence of structure II hydrate formation from methane+ethane mixtures. *Chem. Eng. Sci.* **55**, 1981–1985 (2000).
- Staykova, D. K., Hansen, T., Salamatin, A. N. & Kuhs, W. F. Kinetic diffraction experiments on the formation of porous gas hydrates. *Proc. 4th Int. Conf. Gas Hydrates* (2002).
- Hester, K. & Sloan, E. D. Structure II hydrates from binary structure I simple guests. *Fluid Phase Equilibria* (in the press).
- Davidson, D. W., Handa, Y. P., Ratcliffe, C. I., Tse, J. S. & Powell, B. M. The ability of small molecules to form clathrate hydrates of structure II. *Nature* **311**, 142–143 (1984).
- Ripmeester, J. A., Ratcliffe, C. I. & Powell, B. M. A new clathrate hydrate structure. *Nature* **325**, 135–136 (1987).
- Ballard, A. & Sloan, E. D. The next generation of hydrate prediction: An overview. *Proc. 4th Int. Conf. Gas Hydrates* (2002).
- van der Waals, J. H. & Platteeuw, J. C. Clathrate solutions. *Adv. Chem. Phys.* **2**, 1–58 (1959).
- Mehta, A. P., Hebert, P. B., Cadena, E. R. & Weatherman, J. P. Fulfilling the promise of low-dosage hydrate inhibitors: Journey from academic curiosity to successful field implementation. *SPE Prod. Facil.* **73–79** (2003).
- Kvenvolden, K. A. in *Methane Hydrates: Resources in the near Future?* (Proc. Int. Japan Natl Oil Comp., Chiba City, Japan, 1998).
- Milkov, A. V. & Sassen, R. Resource and economic potential of gas hydrates in the northwestern gulf of Mexico. *Proc. 4th Int. Conf. Gas Hydrates* 111–114 (2002).
- Soloviev, V., Ginsburg, G., Telepnev, E. & Mikhailyuk, Y. *Cryogeothermy and Natural Gas Hydrates of the Arctic Ocean Sediments* (Ministry of Geology USSR, Leningrad, 1987).
- Dallimore, S. R. *et al.* (eds). Scientific Results from the Mallik 2002 Gas Hydrate Production Research Well Program, Mackenzie Delta, Northwest Territories, Canada. *Bull. Geol. Soc. Can.* **585** (in the press).
- Dallimore, S. R., Collett, T. S. & Uchida, T. Summary of Mallik 21–38 Well GSC Bull. 544 1, 1–10 (1999).
- Paull, C. K. & Matsumoto, R. Leg 164 overview. *Proc. ODP Sci. Res.* **164**, 3–10 (2000).
- Moridis, G. *et al.* Numerical simulation studies of gas production scenarios from hydrate accumulations at the Mallik Site, Mackenzie Delta, Canada. *Proc. 4th Int. Conf. Gas Hydrates* 239–244 (2002).
- Gudmundsson, J. & Borrehaug, A. Frozen hydrate for transport of natural gas in *Proc. 2nd Int. Conf. on Natural Gas Hydrates* 415–422 (1996).
- Nakajima, Y., Takaoki, K., Ohgaki, K. & Ota, S. Use of hydrate pellets for transportation of natural gas — II: Proposition of natural gas transportation in form of hydrate pellets in *Proc. 4th Int. Conf. Gas Hydrates* 987–990 (2002).
- Shiota, H. *et al.* Measurement of methane hydrate dissociation for application to natural gas storage and transportation. *Proc. 4th Int. Conf. Gas Hydrates* (2002).
- Gudmundsson, J., Andersson, V., Levik, O. I., Mork, M. & Borrehaug, A. Hydrate technology for capturing stranded gas. *Ann. NY Acad. Sci.* **912**, 403–410.
- Kennett, J. P., Cannariato, G., Hendy, L. L. & Behl, R. J. *Methane Hydrates in Quaternary Climate Change: The Clathrate Gun Hypothesis* (Am. Geophys. Union, Washington DC, 2003).
- Dickens, G. R., Castillo, M. M. & Walker, J. C. G. A blast of gas in the latest paleocene: simulating first-order effects of massive dissociation of ocean methane hydrate. *Geology* **25**, 259–262 (1997).
- Dickens, G., O'Neill, J., Rea, D. & Owen, R. Dissociation of oceanic methane hydrates as a cause of the carbon isotope excursion at the end of the Paleocene. *Paleoceanography* **10**, 965–971 (1995).
- Kaiho, K. *et al.* Latest paleocene benthic foraminiferal extinction and environmental changes at Tawau, New Zealand. *Paleoceanography* **11**, 447–465 (1996).
- Maslin, M. & Thomas, E. The clathrate gun is firing blanks: Evidence from balancing the deglacial global carbon budget. *Geophys. Res. Abstr.* **5**, 12015 (2003).
- Dickens, G. R. A methane trigger for rapid warming? *Science* **299**, 1017 (2003).
- Dillon, W. P. *et al.* in *Gas Hydrates: Relevance to World Margin Stability and Climate Change* (eds Henriet, J.-P. & Mienert, J.) 293–302 (Geol. Soc., London, 1998).
- Paull, C. K. & Dillon, W. P. (eds) *Natural Gas Hydrates: Occurrence, Distribution and Detection* (Am. Geophys. Union, Washington DC, 2001).
- Bishoi, P. R. & Natarajan, V. Formation and decomposition of gas hydrates. *Fluid Phase Equilibria* **117**, 168–177 (1996).
- Sloan, E. D. Jr in *Gas Hydrates: Relevance to World Margin Stability and Climate Change* (eds Henriet, J.-P. & Mienert, J.) 31–40 (Spec. Publ. 137, Geol. Soc., London, 1998).
- Dickens, G. R. Methane oxidation during the Late Paleocene Thermal Maximum. *Bull. Soc. Geol. Fr.* **171**, 37–49 (2000).

Integration of geoscience and engineering in the oil industry — just a dream?

B. Artur Stankiewicz

Shell International E&P, B.V., Volmerlaan 8, 2280 AB Rijswijk, The Netherlands (e-mail: astankiewicz@wanadoo.nl)

The past two decades of the twentieth century have been very 'rocky' for the oil industry, as shown by the overall negative perception of the oil companies by the general public. Fluctuating oil prices, many rounds of staff redundancies, environmental disasters and budget cuts supported the overall image of the oil industry as being 'the technology impaired and environmentally insensible giant'. But advances and positive changes have been quietly happening in most of the oil companies. In the twenty-first century, we will witness the metamorphosis of the oil and gas companies into energy businesses — the era of cleaner and safer oil and gas production, and alternative energy resources such as wind, solar and hydrogen is already underway.

In 1853, Polish pharmacist I. Lukasiewicz perfected the distillation process to produce fuel for oil lamps, creating what was arguably the first market for refined oil. Not long afterwards, the Polish Carpathian Mountains became home to the first 'modern' oil wells and the birthplace of the European oil industry. However, it was in the United States six years later that the oil industry as we know it today burst from the ground, when self-styled Colonel Edwin Drake successfully produced oil from a well in Titusville, Pennsylvania, on 28 August 1859. Although this was certainly news, nobody, with the possible exception of John D. Rockefeller, could have had any idea of the monumental effect that the ability to produce large quantities of oil would have on the twentieth century.

Much has changed in the oil business since the 1850s. Technical advances have led to oil and gas production in almost every part of the world, and in every conceivable environment, from easily accessible sites, termed 'low-hanging fruit', to very hostile conditions such as ultra-deep water (ref. 1 and see the review in this issue by White, page 334). The past 15 to 17 years have seen tremendous changes resulting from industrial globalization and restructuring of the industry. We have witnessed fluctuations in the price of a barrel from US\$10 to US\$35, employment instability, the disappearance of many companies and the birth of new ones; this is a period that deserves its chapter in Yergin's Pulitzer-Prize-winning book *The Prize*². The creation of new super major oil companies (for example, BP, ExxonMobil and TotalFinaElf) is partially driven by a need to become more dynamic and integrated to compete successfully in an increasingly 'tight' crude oil market. Beating the competition requires the development of cutting-edge technology that can be implemented quickly and successfully in the exploration and production of oil and gas. This can be achieved only with the aid of a specialized force of skilled scientists and engineers — the cornerstone of any successful technical business.

Finding the right skills

The pioneers of oil exploitation came from a wide variety of backgrounds and were expected to have a good knowledge of all aspects of exploration and production. The past 60 years have witnessed the development of very specific disciplines, such as petroleum geology, organic geochemistry and reservoir engineering. Specialization, however, can inhibit the development of an integrated approach. There is no question that all of these industry-related disci-

plines do add value to the oil business, but successful exploration or production cannot be based on a one-sided view (Fig. 1). The geophysicist may identify a hydrocarbon trap on the basis of seismic data, but it is worth nothing if the geochemist and basin modeller conclude that the basin lacks the hydrocarbon charge. The petrophysical engineer can describe the rock properties, and the reservoir engineer can estimate the volume of hydrocarbon and model its behaviour, but they both need data on the pressure, volume and temperature of the fluids. The mechanical engineer can design production facilities, but he requires detailed information about the behaviour of the oil and the producing gas-to-oil ratio.

During the past 20 years, there has been an increased focus on successful production, which has led most of the exploration-oriented disciplines to reinvent themselves as production-oriented technologies. For example, research into structural and reservoir geology now focuses on predicting the size of hydrocarbon traps; and geochemistry, traditionally linked entirely to exploration, is now an important production tool. Geoscientists now monitor field production using time-dependent seismic surveys (also known as four-dimensional seismic surveys). And flow assurance has emerged as an important discipline for minimizing the risk of unexpected equipment blockage by hydrates³ (see the review in this issue by Sloan, page 353), scale, wax or asphaltene^{4,5}. Thus, cost-effective development and production of oil from deep-water fields is possible.

Integration synergies

Although niche knowledge is still essential, integration of the various skill sets and expertise is needed to provide reliable solutions to the challenges that the industry faces. An excellent example is the emergence of experts in fluid properties who cover the areas of geology, geochemistry, petrophysics, and reservoir and chemical engineering. These specialists understand the behaviour of petroleum fluids from their generation, through migration and reservoir filling, to production and transportation to topside facilities. Their combination of skills from geosciences and engineering, to subsurface and surface disciplines, provides an integrated view of exploration and production, from source to refinery. What was formerly a typical engineering challenge — the problem of asphaltene deposition in Venezuelan fields — has been solved by geochemists working with flow-assurance engineers⁶. For example, an in-depth investigation of the behaviour of produced fluids, using cutting-edge sam-



Figure 1 Virtual-reality room for state-of-the-art subsurface visualization. Copyright Royal Dutch/Shell.

pling and laboratory techniques, prompted a strategy of controlled flow rates from individual wells, which led to an immediate increase in production. Traditional exploration tools, such as basin modelling, have been used to determine the distribution of fluids in the subsurface with the highest tendency for asphaltene deposition. This information was used to develop a predictive model for near-field exploration. Challenges such as this one were difficult to meet before the integration of data and expertise. In fact, this domain, traditionally that of production chemistry, always lacked subsurface expertise; here we demonstrate the added value of integration and understanding of the whole petroleum system. One of the key elements responsible for expansion of expertise in fluid-properties was the need to develop, safely and profitably, methods of extracting petroleum from deep water (Box 1).

Basic research

Oil companies are in the business of producing fluids, not rocks, and so a full understanding of a fluid's quality, behaviour and properties is



Figure 2 A shoal of fish swimming through an offshore jacket — part of a programme for establishing marine life on a decommissioned offshore plant. Copyright Royal Dutch/Shell.

essential. The need to model compositional grading in large deltaic reservoirs prompted the so-called 'centrifuge experiment', in which fluid behaviour in natural conditions was successfully simulated, and it provided the industry with a model of fluid gradients in many deep-water reservoirs⁷. The impact of that research is far reaching; it can be used to predict changes in fluid properties during the life of a field, and can affect decisions such as well placement and the sizing of the topside facilities. Another example of such a 'return to basics' research area is that of flow assurance: deep-water production challenges have led to extensive research into the behaviour of the heavier components of oil, such as wax and asphaltenes. Various laboratory devices have been developed to investigate the behaviour of fluids and solids under a wide range of pressure and temperature conditions — simulating processes occurring in the reservoir, wellbore and flowlines^{5,8}.

The need to study a fluid's behaviour under laboratory conditions prompted the collection of subsurface fluid samples directly from the reservoir^{8,9}. This is not a trivial undertaking. Consider, for example, the following conditions: the reservoir has a thickness of 30 m, is located 5,000 m below the sea bottom, 100–200 km from a land-based operation in water 1,500 m deep, at a pressure of 800 bar and temperature close to 100 °C. In addition, the well is a 30° deviated hole. Yet obtaining fluids in pristine condition from such a well and transporting them to the onshore laboratory for further analyses has become an almost daily procedure for many new field developments. One of the first and most important measurements of reservoir fluids is its PVT — pressure, volume and temperature. The widespread use of this measure is often not fully appreciated¹¹. Data on the composition, density, viscosity and saturation pressure of the hydrocarbon fluids are as important to the engineers modelling reservoir properties as to those designing production facilities. The flow-assurance experts need such data to ensure safe and profitable production — the error margins for predictions of wax or asphaltene behaviour and the region of hydrate formation of an oil in a subsea flowline must be negligible. Very often the fluids sampled in the exploration or appraisal stage are the only pristine hydrocarbons ever taken from the well — later fields or pay zones are often co-mingled and produced together. The information obtained about the physical fluid properties can be used later in studies of production allocation or reservoir

Box 1

Petroleum exploration and production in the deep-water realm: conquering a challenging environment

Around 75% of estimated remaining oil and gas reserves lie offshore. Over the past two decades, the industry has been forced to move into increasingly deep water to tap these reserves and meet global energy demand (ref. 1 and see the review by White, page 334). But the environment is a challenging one, often with a water depth exceeding 1,000 m, with high waves, freezing temperatures and strong winds. In response, the industry has had to develop rapidly to counter these forces, resulting in major advances in both technology and its application. A significant enabler for these advances has been the integration of diverse disciplines that bring together drilling, production and engineering of facilities with subsurface geology and other sciences. This united approach has not only enabled the oil and gas industry to extend the potential supply of hydrocarbon, but also transformed it into one of the most technologically advanced industries.

One of the earliest moves into deep water (classed as water of 310–1,500 m in depth) was in the Gulf of Mexico in the 1980s. There, the US Mineral Management Service opened up vast deep-water areas for exploration and production. Shell was one of the early winners, with two major field discoveries that set records for water depth when they later produced oil — 400 m for the Bullwinkle field and 800 m for the Auger field.

Today, most major oil companies are successfully producing oil from water depths exceeding 1,000 m and exploring production from depths of more than 1,500 m (termed ultra-deep water) in the offshore regions of West Africa, Brazil, west Shetland, southeast Asia and, of course, the Gulf of Mexico. With the advent of dynamically positioned drill ships, drilling in these water depths is now a matter of routine.

But before tackling the challenges of deep water, and long before a drop of oil can be brought to the surface, the subsurface has to be fully understood (Fig. 2). The combined skills of geoscientists and engineers are needed to reveal the complex underground picture. In the early days of the Gulf of Mexico discoveries, the integration of geological and geophysical techniques resulted in the identification of vast turbidite deposits in ponded basins along the middle continental slope with excellent

reservoir properties. On the basis of the seismic data that have been gathered, stratigraphic methods have been developed to identify these sand-rich facies. Another multidisciplinary task was to map and characterize hydrocarbon traps lying near complex salt bodies that were obscuring the images from subsurface imaging devices. The ability to adapt land-based seismic technology to this more complex environment (involving more sophisticated seismic data acquisition, processing and interpretation techniques) was one of the most important factors in the successful quest to find oil and gas located several kilometres below the seabed. Since then, many of the geophysical techniques developed have been adopted for detecting drilling hazards at the seabed where production facilities will be placed, and for routing oil and gas transportation pipelines around environmentally sensitive sites, such as colonies of unusual chemosynthetic organisms and groups of animals that thrive near naturally occurring oil seeps. Furthermore, mapping of these seeps, together with geochemical typing, was one of the tools that provided data that enabled detailed basin modelling and later recognition of the prolific Mesozoic petroleum system in the Gulf of Mexico and the Tertiary-sourced oils of the Niger Delta.

With deep-water wells costing from US\$20 million to more than US\$100 million, oil production rates of 5,000 to more than 10,000 barrels per day are essential, and every piece of information is critical to achieve this goal. Integrated knowledge and skills, combined with technologies that allow bigger perforations (and hence, greater flow of oil into the well), longer completion intervals and wider production tubing now make the high-rate wells needed for the deep-water environment the norm rather than the exception.

But this improved performance has not come entirely from the efforts of the oil and gas industry. Effective ties between civil engineers and oceanographers/meteorologists have been essential in the design, installation and operation of safe and reliable production platforms. Designing fixed and floating structures requires a sound understanding of expected wave heights, the influence of eddy currents, storm intensities and frequencies — all of which vary locally and seasonally. Sophisticated sub-sea engineering equipment is used to install and operate well-heads, manifolds and pipelines that sit on the seafloor in water routinely exceeding a depth of 1,000 m.

Technologies ‘borrowed’ from other industries have found their home in the oil and gas industries, such as virtual-reality screens to view, three-dimensionally, subsurface data, or computer axial tomography (CAT) scans used by the medical world to ‘see through’ rocks. Satellites are also being used to operate unmanned facilities, to keep extremely heavy drilling rigs in position (they do not vary in position by more than 1 m, even in rough seas) and to view the world in pin-sharp detail.

In conclusion, the successful development of deep-water petroleum resources has required collaboration between a large number of skilled experts from several disciplines. To continue this success story, it is very important that our society values and supports education in the earth sciences and petroleum engineering. For their part, oil companies need to continue encouraging teamwork and supporting innovation, as well as research efforts that facilitate the sustainable development of affordable energy supplies. Finally, the deep-water story contradicts a commonly held picture of the petroleum industry — that it is an old-fashioned, ‘sunset’ endeavour that is not environmentally conscious. In truth, this industry is at the forefront of developing and applying cutting-edge technologies, effectively and responsibly.



Box 1 Figure 1 The Ram/Powl tension-leg platform (TLP) — permanently anchored in water 980 m deep in the Gulf of Mexico. Copyright Royal Dutch/Shell.

compartmentalization — studies that require the expertise of production geochemists. Moreover, knowledge of the type and quality of fluid samples needed for analysis are critical not only to the ultimate success of the project but also for controlling the capital and operational expenditure. The fluid-properties specialist is at the forefront of sample acquisition. No longer do exploration and development teams send scientists plastic jars and say, “tell me what this fluid is”; now, thanks to the quality control provided by fluid-properties specialists, exploration and development teams are expanding their knowledge of a fluid’s behaviour in the subsurface and in production facilities.

Building the future in a sustainable way

The change of focus in the oil industry has led to an increased emphasis on five main areas: hydrocarbon production, basic science and technology, the integration of disciplines and expertise, environmental and safety aspects, and sustainable development (Fig. 2). In the past 25 years, oil companies have started to explore and produce oil from fields in economically and technologically challenging areas such as deep-water environments (Box 1) (ref. 1 and see the review by White, page 334), the Arctic and high-pressure/high-temperature reservoirs. Although enormous reserves of crude oil and gas remain in relatively accessible settings (that is, onshore or in very shallow water, up to 50 m deep), most international oil companies are exploring ways to find and produce oil and gas in more difficult environments. Unconventional resources such as oil shales, tar sands, tight gas and coal-bed methane are receiving increased interest. Reserves of these hydrocarbon resources exceed those of conventional sources, but they can be more challenging to unlock and will receive increased attention in the years to come. Companies must also safeguard the environment and consider issues related to carbon dioxide emissions, the rising level of greenhouse gases and global warming (see the review by Berner, page 323). The Athabasca Oil Sands Project in Canada shows how unconventional reserves can be exploited with the incorporation of sustainable development strategies at every stage of the project. The production of synthetic crude oil started in 2003 using a process of upgrading heavy crude by the addition of hydrogen. There are also world-wide commitments to reducing greenhouse gas emissions by up to 50% by 2010. Many oil companies are working to reduce carbon dioxide emission by implementing new technologies to eliminate continuous flaring (burning associated gas at topside facilities).

The increasing need for sustainable development has galvanized creativity and out-of-the-box thinking. Companies are pursuing alternative forms of energy, such as wind, solar or hydrogen power, as well as seeking traditional ‘clean’ energy sources such as natural gas. These new avenues are the future for oil companies... or perhaps we should call them energy companies? The increasing focus on the environment has led to programmes that allow employees to contribute directly to research on ecology, biodiversity, conservation, community issues and climate change (Fig. 2). Engineers and geoscientists can expand their horizons beyond oil production. What, you ask, has this to do with integration and technology in the oil industry? Nothing perhaps, but the image of a production engineer monitoring the behaviour and migration of South African penguins reflects a new approach, unheard of even ten years ago. The petroleum indus-

try is changing, and the breaking down of disciplinary boundaries and ‘thinking beyond oil’ are no longer fantasy but an exciting reality.

Personal perspective on technology and work in an oil company

Despite my training as a geologist and organic geochemist, I really did not know what to expect when I first joined the oil industry, working for one of the major players. My postdoctoral research in Bristol on the preservation of animal cuticles and human remains in the fossil record impressed my future employer more than my PhD thesis on coal and kerogen, and this played an important part in my decision to join ‘the business’ and leave ‘academia’. It took 12 months for me to realize that what they wanted was my ability to work ‘outside the box’ and tackle new avenues of research using ‘traditional’ (geochemical) knowledge. By this time I was working as a geochemist in the engineering-oriented flow-assurance team. Three years later, not forsaking my geochemical background, I had embraced the way of the engineer, had learned engineer-speak and was working on engineering projects. I have learned from personal experience that global knowledge sharing and the synergies developed from the integration of seemingly unrelated skill sets is vital to producing the thinking that is necessary to tackle today’s exploration and production challenges. The effort involved in tackling novel challenges is often daunting and requires me to step outside my ‘comfort zone’. But the personal and intellectual rewards are worth the effort — the possibilities of expanding horizons are almost unlimited. The oil industry is still a cyclical business: funding for research and development has its ups and downs. R&D funding can be increased one year and cut the next. The process varies between companies, and experiences differ. Yet, regardless of ‘thin’ or ‘fat’ years, cutting-edge technology is being developed and implemented, and it is fun to be instrumental in this process. □

doi:10.1038/nature02136

1. Leffler, W. L., Pattarozzi, R. & Sterling, G. *Deepwater Petroleum Exploration & Production. A Nontechnical Guide* (PennWell, Tulsa, 2003).
2. Yergin, D. *The Prize: The Epic Quest for Oil, Money & Power* (Simon & Schuster, London, 1991).
3. Mehta, A., Walsh, J. & Lorimer, S. Hydrate challenges in deep water production and operation. *Ann. NY Acad. Sci.* **912**, 361–373 (2000).
4. Hammami, A. & Raines, M. A. Paraffin deposition from crude oils: Comparison of laboratory results with field data. *SPE J.* **4**, 9–18 (1999).
5. Karan, K., Hammami, A., Flannery, M. & Stankiewicz, A. Evaluation of asphaltene instability and chemical control during production of live oils. *Petrol. Sci. Tech.* **21**, 629–645 (2003).
6. Stankiewicz, B. A. *et al.* in *Abstr. 20th Int. Meet. Org. Geochem.* Vol. I, 121–122 (2001).
7. Ratulowski, J., Fuex, A. N., Westrich, J. T. & Sieler, J. J. Theoretical and Experimental Investigation of Isothermal Compositional Grading. *SPE Reserv. Eval. Eng.* June, 168–175 (2003).
8. Hammami, A., Phelps, C. H., Monger-McClure, T. & Little, T. M. Asphaltene precipitation from live oils: an experimental investigation of the onset conditions and reversibility. *Energy Fuels* **14**, 14–18 (2000).
9. Jamaluddin, A. K. M., Ross, B., Calder, D., Brown, J. & Hashem, M. Single-phase bottomhole sampling technology. *J. Can. Petrol. Technol.* **41**, 25–30 (2002).
10. Kool, H., Azari, M. & Soliman, M. Y. Sampling challenges in gas condensate reservoirs. *Petromi* **28**, 26–30 (2002).
11. Macleod, G. & Meulbroek, P. *Integration of Organic Geochemistry and Pvt Studies in Petroleum Exploration and Production*. (Oxford Univ. Press, Oxford, 2004).

Acknowledgements I would like to thank J. Westrich, D. Briggs, M. Nosiara and R. Hilliard for review and valuable suggestions during work on this Commentary. Box 1, apart from the knowledge gathered during my tenure at Shell, is based on interviews and information provided by the following colleagues: S. Lorimer, F. Pattee, J. Siler, L. Morgenthaler, J. Westrich and A. Kornacki. The comments are the views of the author and not necessarily those of Shell.

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Netherlands/ Italy/ Spain/

Portugal/ Belgium:

Evelina Rubio Hakansson (4973)
Scandinavia: Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager:

Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harajukutamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Europe goes back to basics

The arguments in support of a single 'European research area' received a fresh airing last month at a meeting in Dublin. Participants heard how the United States is dominating science, with a disproportionate number of highly cited papers — and Nobel prizewinners — now coming from US labs. One factor seemed particularly pertinent — many of the lauded authors and laureates are actually Europeans who have moved across the Atlantic.

A recurring theme at the meeting was the need for more European-led basic research, says Enric Banda, secretary-general of the European Science Foundation (ESF). Currently, funding for basic research varies from country to country within the European Union. And the European Commission's funding mechanism, the Sixth Framework Programme, emphasizes applied research and collaborative projects.

The Dublin meeting, convened by the European Life Sciences Forum and EuroScience, tried to address the issue. One way to combat the brain drain would involve a central European Research Council that would fund basic research across all member countries. Another would be to hold member nations to the commission's goal of increasing each country's research-and-development funding to 3% by 2010.

But bringing about such changes is beyond scientists' control. A more realistic approach is to pursue incremental changes from the ground up. The European Young Investigator Awards, which will provide up to €250,000 (US\$294,000) a year for five years and will be chosen by the ESF next month, represent one such example. But they are just a first step — the earliest possible date that a European Research Council could form is 2007. In the meantime, Europeans can but hope that the Young Investigator Awards lead to high-impact research — and a stop-gap measure against the brain drain.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

Rhône-Alpes heads
for the summit

p366

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Twin peaks Rhône-Alpes

The Rhône-Alpes region of France has a rich cultural history, a reputation for innovation and creativity — and great skiing. But researchers are now just as likely to come for the quality of the research environment as the ski slopes. The main cities of Grenoble, gateway to the Alps and a hotbed of technological development, and the more cosmopolitan Lyons, proud of its long history in biomedicine, have always jealously protected their individuality. But now, in research at least, they are realizing that strength comes through synergy rather than rivalry.

A proactive regional council has doubled its financial support for research over the past four years — contributing E43 million (US\$50 million) in 2003. Now a combination of public and private targeted investment is bringing coherence to the region's research infrastructure and, investors hope, increasing its international visibility.

Lyons' biomedical strengths include one of the biggest concentrations of hospitals in Europe: some 6,000 beds provide plenty of clinical-research opportunities for the area's large academic community as well as for researchers from industry and other institutions. The city is home, for example, to the

World Health Organization's International Agency for Research on Cancer (IARC), to the human-vaccine company Aventis Pasteur, and to the major *in vitro* medical-diagnostics company bioMérieux. In fact, more than a third of jobs in the global vaccine and immunovirology industry are located in Lyons.

And Grenoble is advancing its reputation in engineering in a micro- and nanotechnology project that is one of the region's most ambitious to date. A E400-million innovation and educational centre called Minatéc will house 4,000 people, including 1,000 students, by 2005.

Grenoble also houses three major international laboratories: the European Synchrotron Radiation Facility, a third-generation synchrotron light source; the Laue-Langevin Institute, the world's most intense neutron source; and an outstation of the European Molecular Biology Laboratory. With CERN, the European particle-physics lab, only 90 minutes away in Geneva, the region's labs have access to a unique concentration of large-scale investigative tools for both physical and life sciences.

But until recently the multitude of research partners across the region tended to work independently, forging alliances that have never grown big enough to compete nationally and internationally. "There has been too much individualism," says Serge Petit, director of Idéalp' Pharma, a drug-discovery company based near Lyons. "It is a luxury that we can no longer permit ourselves."

PLANTING POLES

The need for interdependence has led to the creation of three research 'poles'. These will link the region's life-science and technology labs together to reduce duplication and focus research efforts on high-throughput technologies, in order to study functional genomics and develop clinical applications, particularly for cancer.

Clear view: a researcher gets to grips with the imaging facility at the functional-genomics network Genopole.



GETTY



STUDIO DE LA REVIRÉE

Lighting the way: the European Synchrotron Radiation Facility — one element of Grenoble's high-tech mix.

Genopole Rhône-Alpes, a functional-genomics network linking nearly 60 laboratories, was launched in 2000. The two others are still being built up: Canceropole as a network of cancer R&D expertise and Nanobio as a programme bringing nanoscience into biotechnology. All three poles will use public-private partnerships to provide funding for priority research projects and improve access to large-scale instruments and technology.

Genopole also supports an exploratory genomics centre, split between Lyons and Grenoble, to develop research and applications in bioinformatics, one of the area's particular strengths. At the other end of the functional-genomics chain, the Lyons-based Animage facility will provide high-throughput imagery of small animals for functional exploration of genes in the whole organism.

Regional funders expect cancer-related research to lead the economic development and clinical applications generated by Genopole's activities. The Canceropole project will receive some E60 million during the next four years from national and regional funds, with industrial partnerships currently under negotiation. Canceropole's director, Jean-Yves Bonnefoy, hopes that the funding will help to create a 'machine' to spin off companies — he aims to launch two per year within the next four years, he says.

And the other end of the pipeline is not being forgotten. "We are reviewing the way we teach cancer at university. There will now be specific oncology training at engineer level," says Charles Dumontet, life-sciences vice-president at Lyons' Claude Bernard University. "There is now a very strong interface between clinical and basic research in Lyons. We're counting on this to attract new researchers," he adds.

Protein'eXpert, a spin-off company based in Grenoble, illustrates how the new regional dynamic is helping to develop such partnerships. Founded in 2000 by two young researchers from Grenoble's Institute of Structural Biology (IBS), the company emphasizes

high-throughput therapeutic and target-protein discovery, production and engineering services. It has grown fast to 25 employees and hopes to have 50 within a couple of years.

Protein'eXpert is working with Genopole's Grenoble-based proteomics centre — which provides high-throughput protein analysis — and has made strong links with the Partnership for Structural Biology, a new initiative of the three international labs in Grenoble and the IBS.

LIFE SUPPORT

Life-sciences spin-off companies are well served in Rhône-Alpes. Two regional incubators, Crealys and GRAIN, already support some 25 projects every year. Biotech businesses coming out of incubation will soon benefit from two new projects: Bioparc in Lyons and Biopolis in Grenoble. These will provide more office space and R&D facilities to complement the well-developed science parks that already exist. And with Bioparc and Biopolis each situated next to a university hospital and health complex, intellectual links with academic labs should flourish.

At least one of the companies joining Biopolis is likely to be working in neuroscience. In 2006 a new clinical neurology laboratory funded by Joseph Fourier University will open next door. Its researchers will work in tandem with the biomedical neuroscience institute in Lyons.

Industry sees the potential that lies in using the synergy between Lyons' biomedical and Grenoble's technological excellence. A E10-million investment by Lyons-based bioMérieux in a site opposite the French atomic-energy commission in Grenoble will provide a permanent home for its fast-growing biochip spin-off Apibio, which emerged from a government laboratory. The company may also develop a molecular-biology pole there in the future. "There is great complementarity between Grenoble and Lyons," says Christophe Mérieux, vice-president of bioMérieux. "The region has huge potential."

These recent investments and collaborations have raised the profile of Rhône-Alpes. Time will tell if sustained investment will be forthcoming to consolidate these new alliances into twin peaks of excellence in Lyons and Grenoble and build up a reputation in life sciences and nanotechnology as high, and as solid, as its mountains.

Sally Goodman is a freelance writer based in Paris.